Supplemental Figures

Figure S1. Schematic of bacterial one-hybrid zinc finger binding site selections
The DNA-binding specificities for candidate "test" zinc fingers were determined by selection of binding sites from a random DNA library as previously reported and described (see **Supplemental Methods 1d**). (a) The figure shows an example of a F2 candidate and the 9 random bases it may interact with (depicted as rainbow-colored bases). F3 candidates are characterized in the analogous way. The complete sequence of an example candidate is listed below the scheme, where the sequence of the test helix is shown in green. (b) Schematic of zinc finger binding site selection. (Top) Test zinc fingers are expressed as a 3-fingered protein-direct fusion to the omega subunit of RNA polymerase. Binding sites are selected from a 28 base pair region of random DNA sequences upstream of the promoter that drives the reporter genes, HIS3 and URA3. (Bottom) Two plasmids, the zinc finger expression vector and the random library reporter vector, are transformed into the bacterial strain. Double transformants are plated on selective media. DNA is recovered from bacteria that survive the selection and the binding site region of library reporter vector is recovered. After motif finding, selected binding sites are shown as a sequence logo. For simplicity, these have been trimmed to the 3 bases selected by the test finger for display in all other figures. (Next page.)

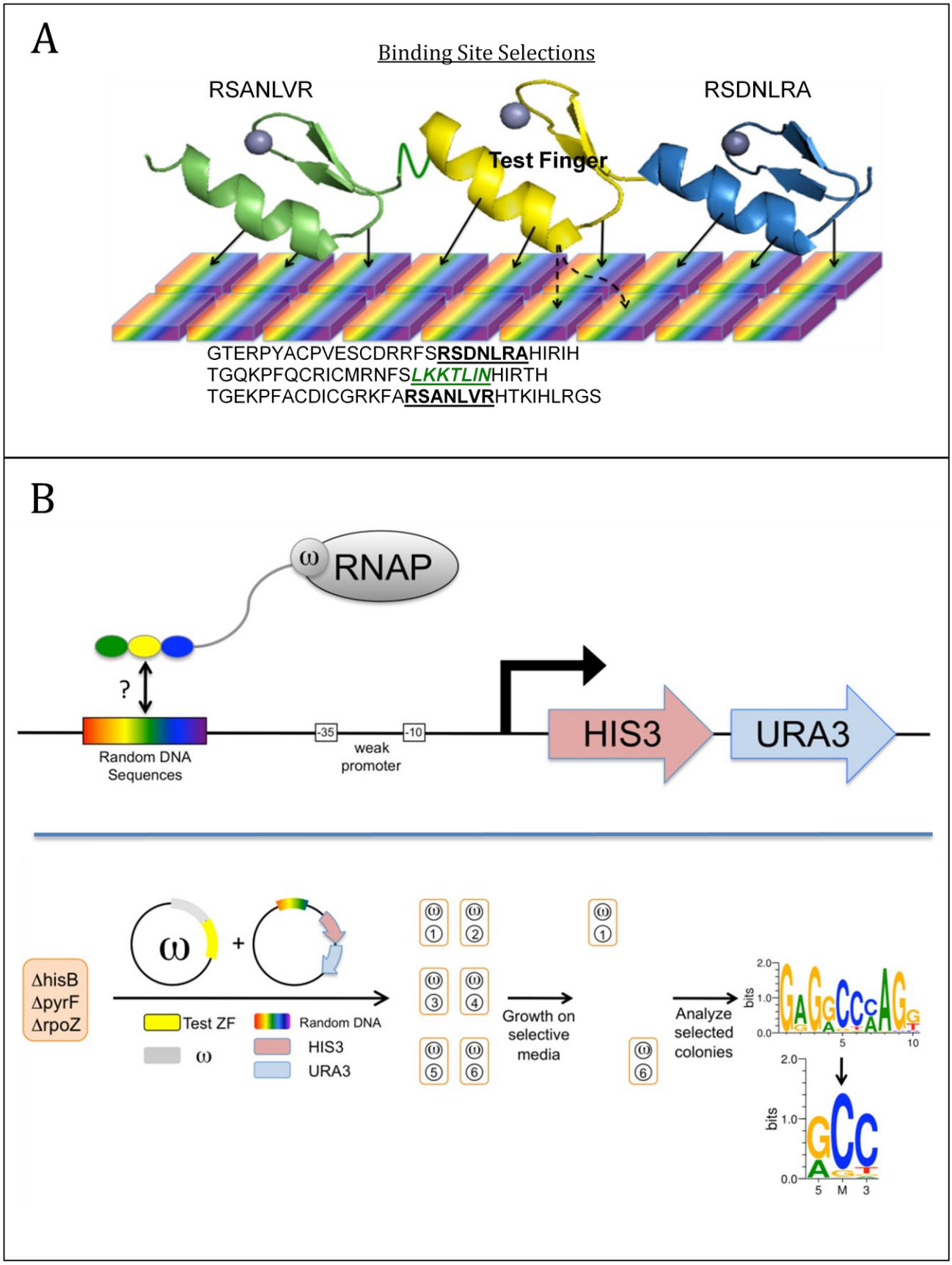

A
Binding Site Selections
RSANLVR
RSDNLRA
Test Finger
GTERPYACPVESCDRRFSRSDNLRAHIRIH
TGQKPFQCRICMRNFSLKKTLINHIRTH
TGEKPFACDICGRKFARSANLVRHTKIHLRGS
B
ω
RNAP
?
Random DNA Sequences
-35
-10
weak promoter
HIS3
URA3
ΔhisB
ΔpyrF
ΔrpoZ
Test ZF
ω
Random DNA
HIS3
URA3
Growth on selective media
Analyze selected colonies
bits
2.0
1.0
0.0
5
10
M
3

Figure S2. Agreement between high and low stringency protein selections. For each 3bp DNA target, we compute the weighted fraction of core sequences found in high stringency that are also found in low stringency in the F2 (left) and F3 (right) positions (see **Supplemental Methods 2b**). We visualize these weighted fractions across the 64 possible targets via boxplots, as in Figure 2b. For most targets, this weighted fraction is close to 1, with relatively small variance across the targets, but with a few notable outliers. Thus, for most targets in both F2 and F3, a large fraction of the frequency-weighted population of core sequences observed at high stringency is also observed at low stringency. Overall, the overlap between the independent high and low stringency selections shows the high quality and reproducibility of the data.

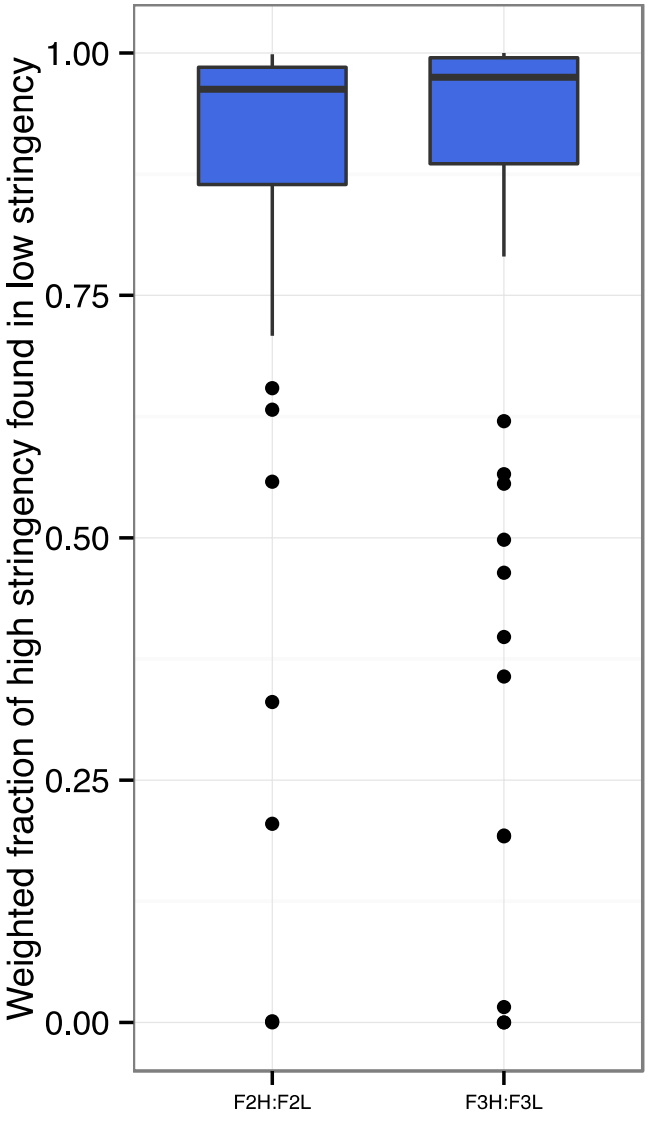


Figure S3. The number of triplets bound by a core sequence is correlated with its per-DNA-target frequency of binding. For each 3bp DNA target, we give the Spearman rank correlation coefficient between the frequency with which each core sequence was observed in the combined F2+F3 protein selections with the number of DNA targets bound by that core sequence across the dataset. For all targets, there is positive correlation between number of targets bound by a core sequence and its frequency of binding within an individual target. *P*-values obtained via a permutation test show all but three of these correlations to be statistically significant (FDR < 0.05 using the Benjamini-Hochberg procedure).

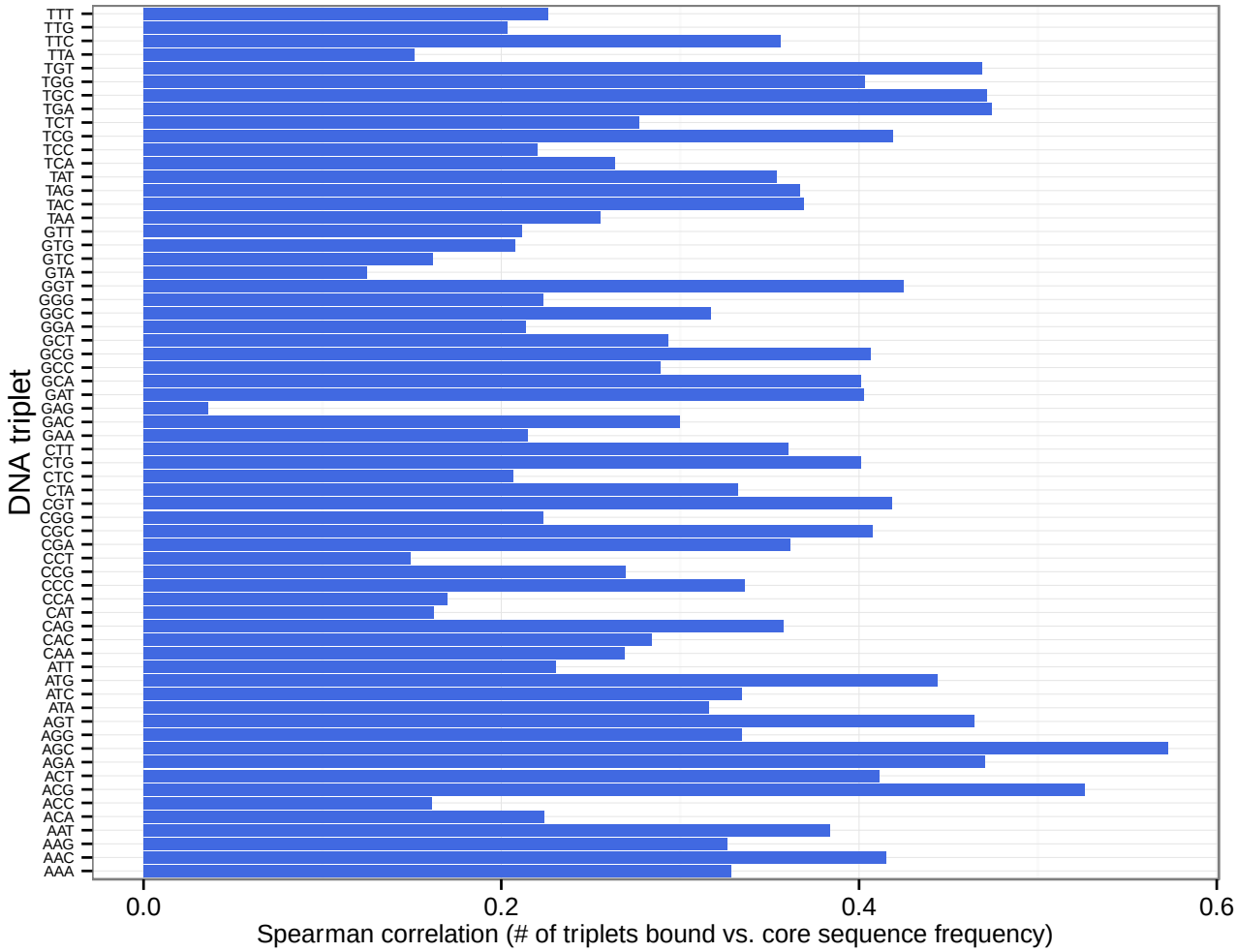


Figure S4. Similar 3bp DNA targets tend to be bound by similar core sequences. We show the frequency weighted overlap of the core sequences binding each pair of targets (see **Supplemental Methods 2b**) in the F3 protein selections as a heat map. The darkest shade of blue represents a score of 1 (complete overlap) and white represents a score of 0 (no overlap). As is apparent by the various patterns in the heat map (diagonal stripes and triangles along the upper diagonal), the target pairs showing highest degree of similarity tend to differ in only one base position.

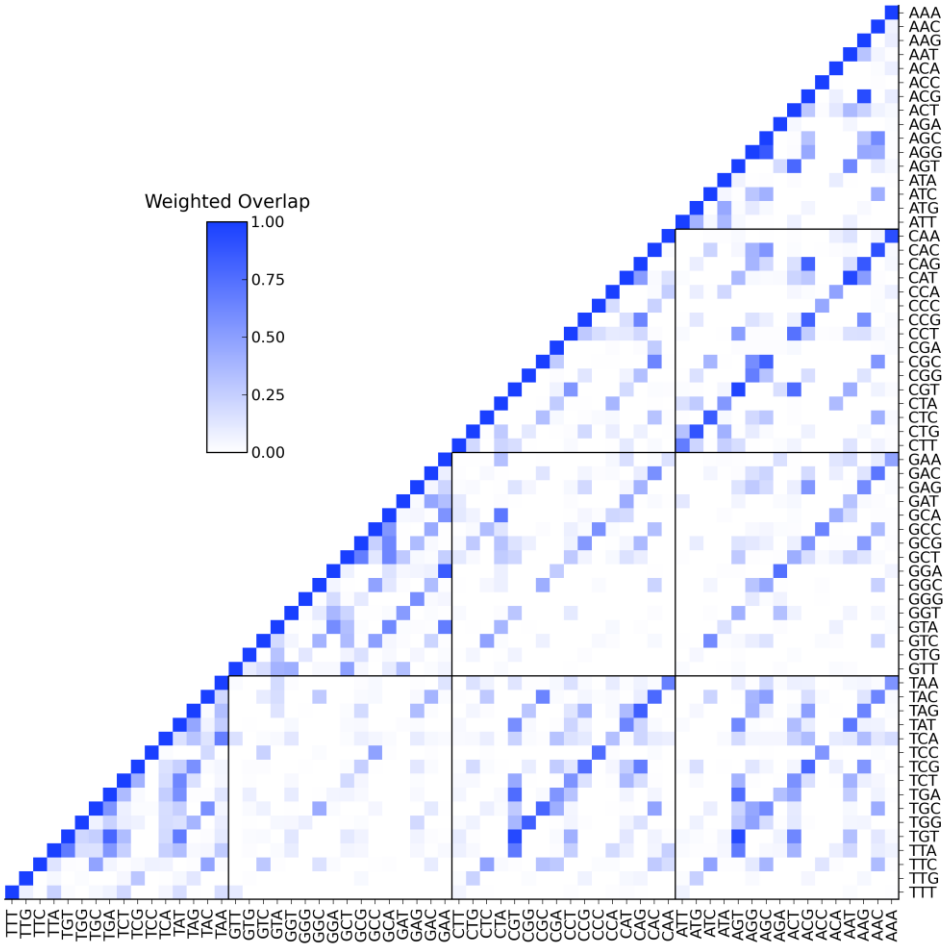


Figure S5. Core sequences recovered from both high and low stringency protein selections tend to have similar binding profiles. For each of the core sequences that were selected at both high and low stringency (1,549 for F2 and 1,482 for F3), we compared their independently inferred binding profiles (computed as described in **Methods**) using the cosine similarity measure and plotted the total frequency of core sequences that fell into discrete similarity bins (**Supplemental Methods 2b**). The majority of the density for this distribution lies in the high-cosine similarity region. Thus, the independently performed high and low stringency selections show good agreement with respect to the binding profiles computed for core sequences.

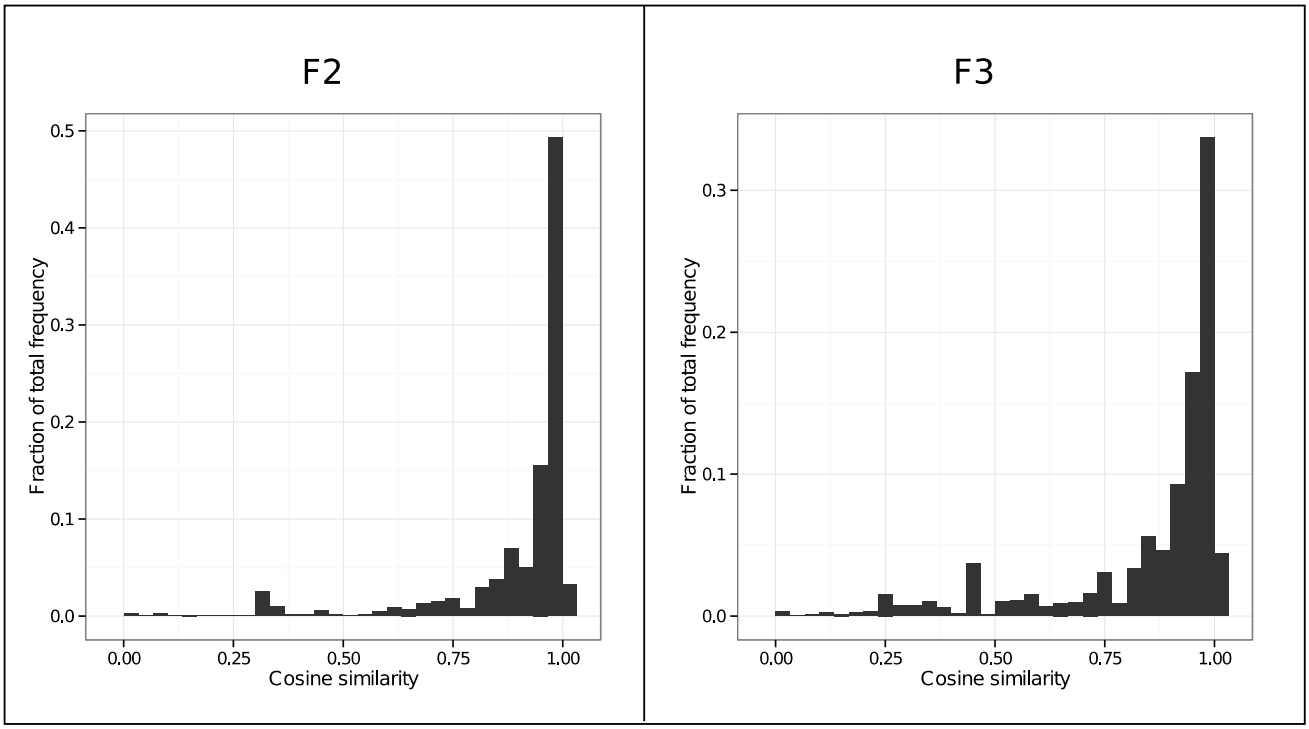
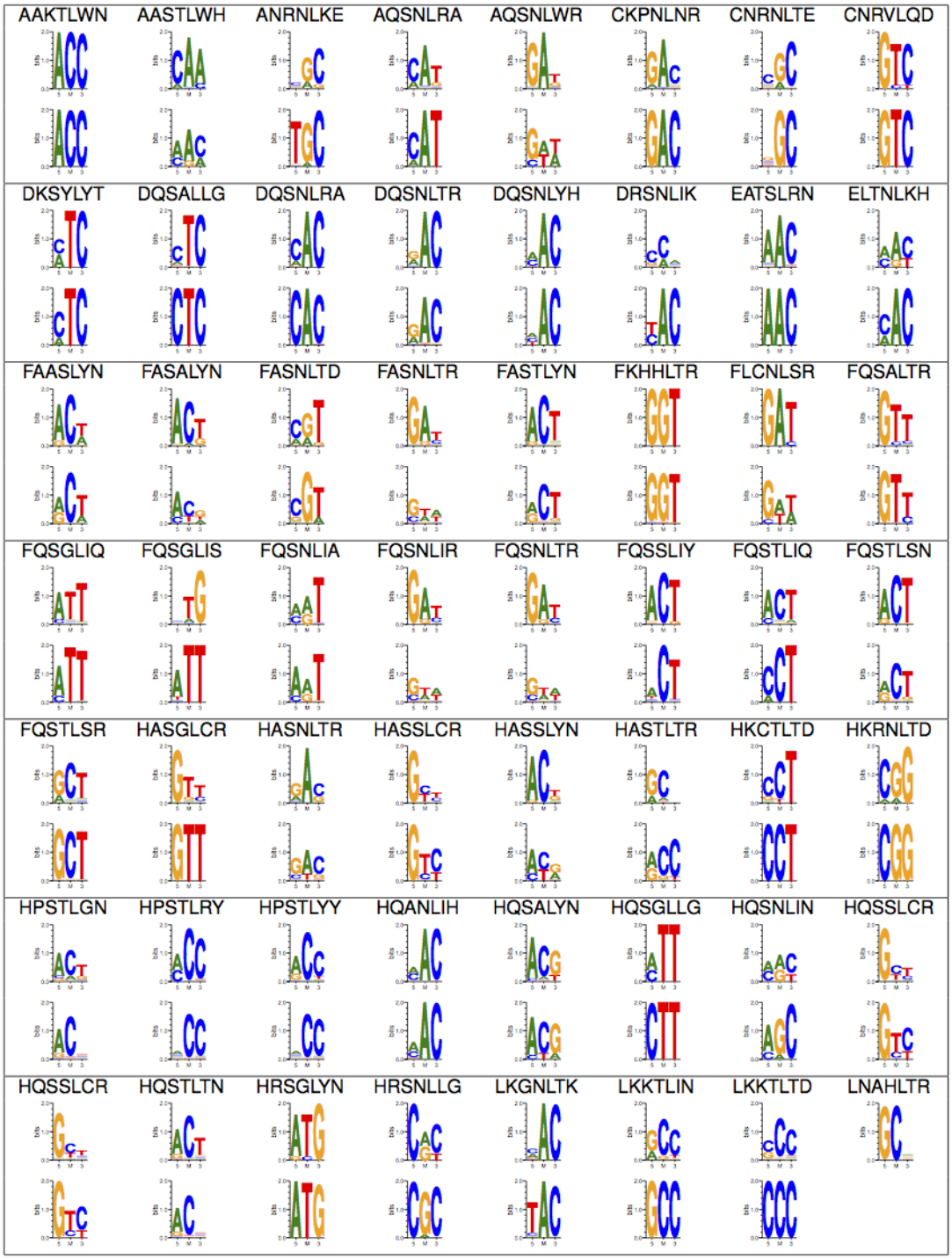
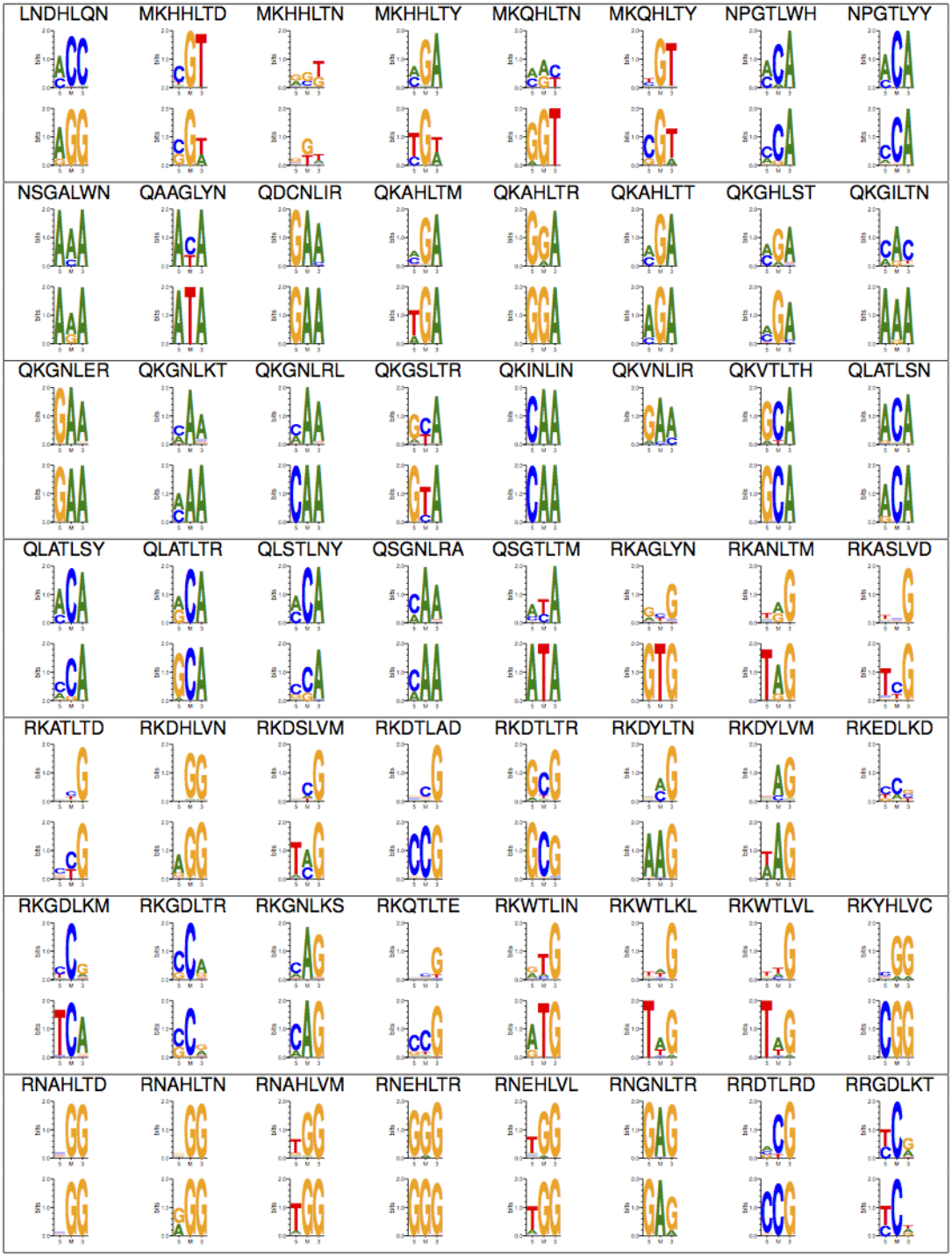


Figure S6. DNA-binding specificities inferred via lookup based upon protein selections show excellent correspondence with experimentally determined DNA-binding specificities. For each of 166 helices from which DNA-binding specificities were successfully obtained in the F2 positional context, we give a paired entry with an experimentally determined logo (top) and a logo computationally inferred from a helix's binding profile in the F2 selections and the lookup procedure, as described in **Methods** (bottom). If the core sequence for a protein is not present in any of the F2 selections, there is no predicted specificity and this is indicated by the absence of a logo. Approximately 83% of the 498 (=166 x 3) paired nucleotide frequency vectors (experimental vs. inferred) are in good agreement (Pearson correlation coefficient >= 0.5) and >67% of the logos show agreement in all three base positions. (Next 3 pages.)


AAKTLWN
AASTLWH
ANRNLKE
AQSNLRA
AQSNLWR
CKPNLNR
CNRNLTE
CNRVLQD
DKSYLYT
DQSALLG
DQSNLRA
DQSNLTR
DQSNLYH
DRSNLIK
EATSLRN
ELTNLKH
FAASLYN
FASALYN
FASNLTD
FASNLTR
FASTLYN
FKHHLTR
FLCNLSR
FQSALTR
FQSGLIQ
FQSGLIS
FQSNLIA
FQSNLIR
FQSNLTR
FQSSLIY
FQSTLIQ
FQSTLSN
FQSTLSR
HASGLCR
HASNLTR
HASSLCR
HASSLYN
HASTLTR
HKCTLTD
HKRNLTD
HPSTLGN
HPSTLRY
HPSTLYY
HQANLIH
HQSALYN
HQSGLLG
HQSNLIN
HQSSLCR
HQSSLCR
HQSTLTN
HRSGLYN
HRSNLLG
LKGNLTK
LKKTLIN
LKKTLTD
LNAHLTR


LNDHLQN
MKHHLTD
MKHHLTN
MKHHLTY
MKQHLTN
MKQHLTY
NPGTLWH
NPGTLYY
NSGALWN
QAAGLYN
QDCNLIR
QKAHLTM
QKAHLTR
QKAHLTT
QKGHLST
QKGILTN
QKGNLER
QKGNLKT
QKGNLRL
QKGSLTR
QKINLIN
QKVNLIR
QKVTLTH
QLATLSN
QLATLSY
QLATLTR
QLSTLNY
QSGNLRA
QSGTLTM
RKAGLYN
RKANLTM
RKASLVD
RKATLTD
RKDHLVN
RKDSLVM
RKDTLAD
RKDTLTR
RKDYLTN
RKDYLVM
RKEDLKD
RKGDLKM
RKGDLTR
RKGNLKS
RKQTLTE
RKWTLIN
RKWTLKL
RKWTLVL
RKYHLVC
RNAHLTD
RNAHLTN
RNAHLVM
RNEHLTR
RNEHLVL
RNGNLTR
RRDTLRD
RRGDLKT

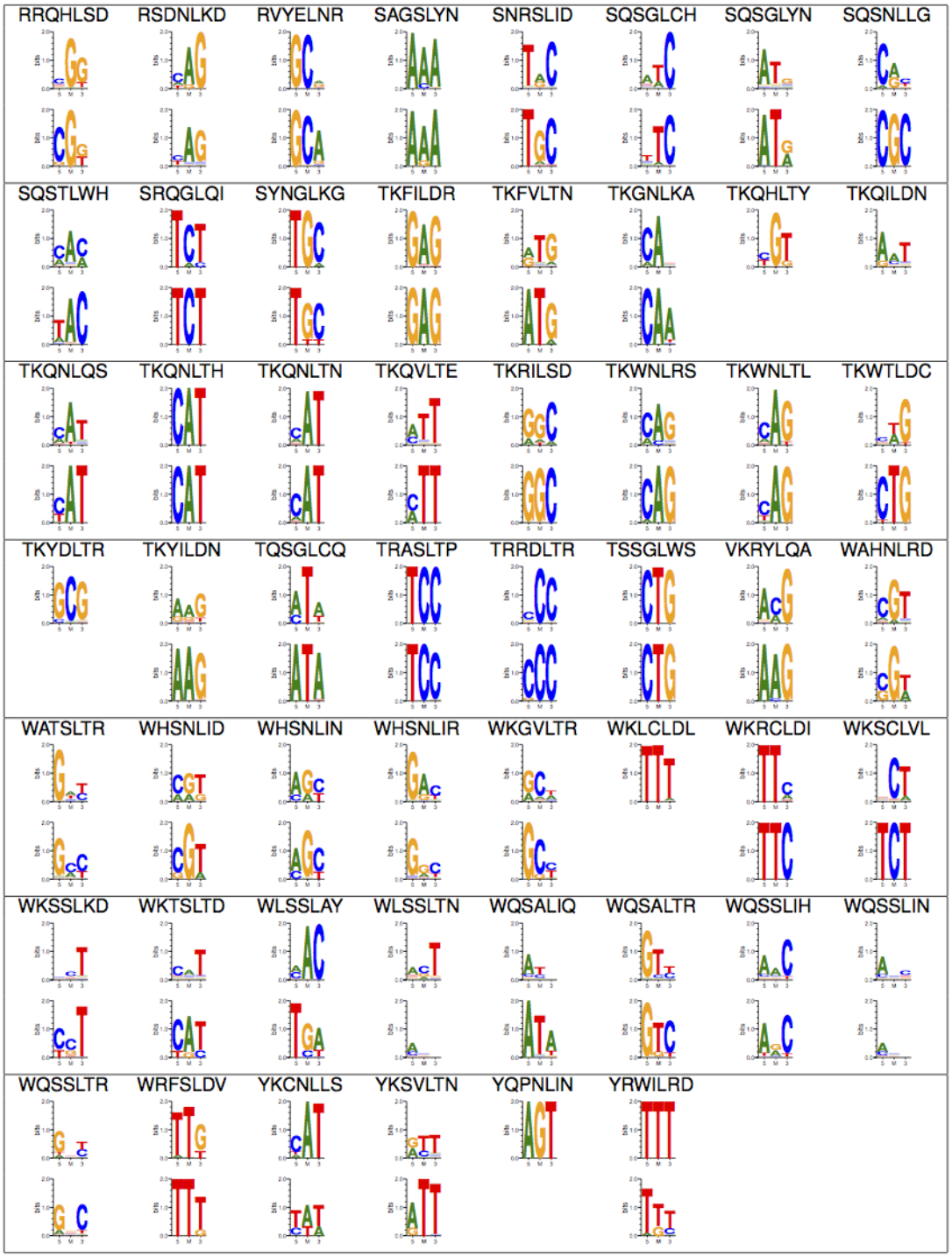

RRQHLSD
RSDNLKD
RVYELNR
SAGSLYN
SNRSLID
SQSGLCH
SQSGLYN
SQSNLLG
SQSTLWH
SRQGLQI
SYNGLKG
TKFILDR
TKFVLTN
TKGNLKA
TKQHLTY
TKQILDN
TKQNLQS
TKQNLTH
TKQNLTN
TKQVLTE
TKRILSD
TKWNLRS
TKWNLTL
TKWTLDC
TKYDLTR
TKYILDN
TQSGLCQ
TRASLTP
TRRDLTR
TSSGLWS
VKRYLQA
WAHNLRD
WATSLTR
WHSNLID
WHSNLIN
WHSNLIR
WKGVLTR
WKLCLDL
WKRCLDI
WKSCLVL
WKSSLKD
WKTSLTD
WLSSLAY
WLSSLTN
WQSALIQ
WQSALTR
WQSSLIH
WQSSLIN
WQSSLTR
WRFSLDV
YKCNLLS
YKSVLTN
YQPNLIN
YRWILRD

Figure S7. A diverse set of helices are found to specify each 3bp target. For each 3bp target (organized as in Figure 4 in the main text), we visualize via sequence logos all the helices observed in our selections that are computationally predicted to specify it via lookup of binding profiles (described in **Methods**). Helices from F2 selections are shown on the left, and from F3 selections are shown on the right. Sequence logos show all 6 variable amino acid positions and the fixed Leucine in position 4 of the helix. Within each logo, the frequency with which each helix is observed across all selections (i.e., the sum of the per-target frequencies computed in **Supplemental Methods 2a**) is used to weight that helix's contribution to the logo.

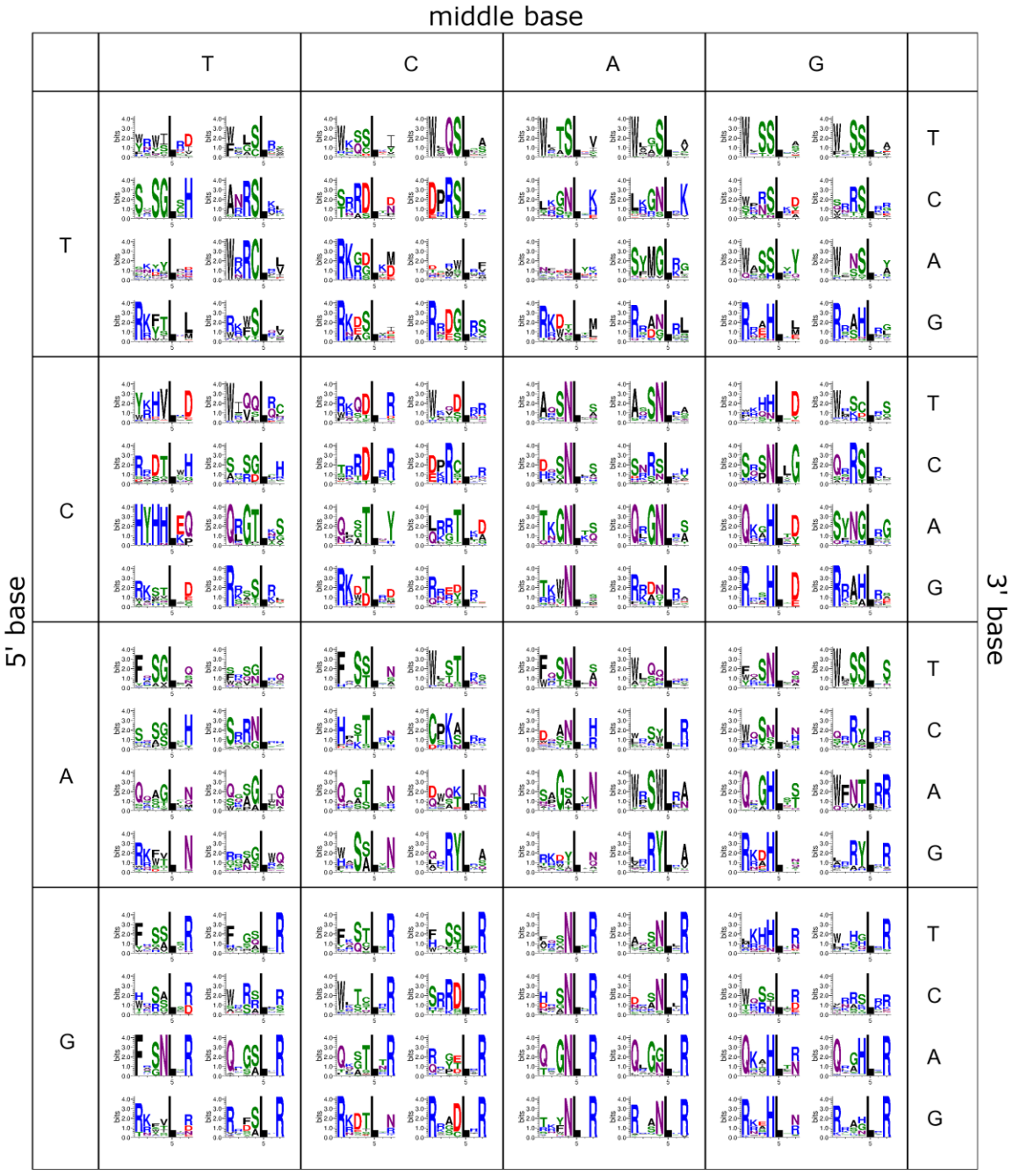
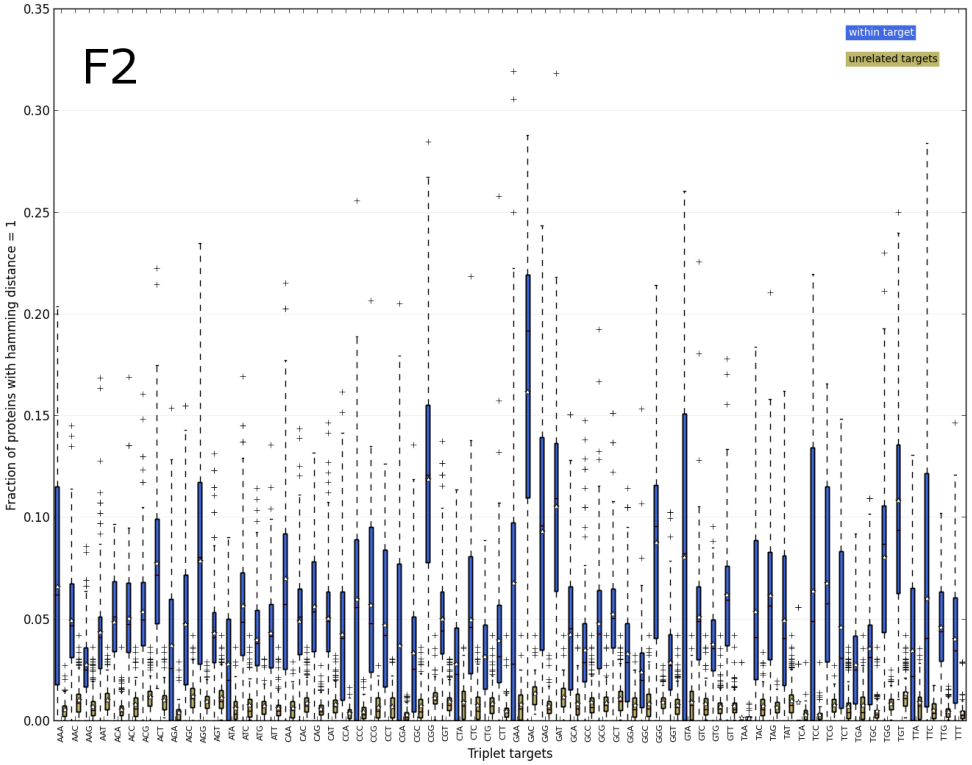


Figure S8. Similar core sequences are found specifying each 3bp target. For each 3bp target, for each core sequence, *C*, assigned to it via lookup using either F2 (top, (a)) or F3 (bottom, (b)), we computed the fraction of other core sequences assigned to that target that were identical to *C* in 3 of the 4 core positions of the recognition helix, and displayed this as a boxplot (shown in blue). For each target, *T*, for each core sequence assigned to it, we also computed this fraction across targets that do not share a common base with *T* in any of the three positions, and displayed this as a boxplot (shown in brown). A lower fraction of core sequences identical in 3 of 4 positions were found when looking across unrelated targets (124 out of 128 F2 and F3 targets were statistically significant at FDR < 0.05, with *p*-values obtained via the Mann-Whitney U-test and FDR correction via the Benjamini-Hochberg procedure). (Next page.)

A


F2
within target
unrelated targets
Fraction of proteins with hamming distance = 1
Triplet targets

B

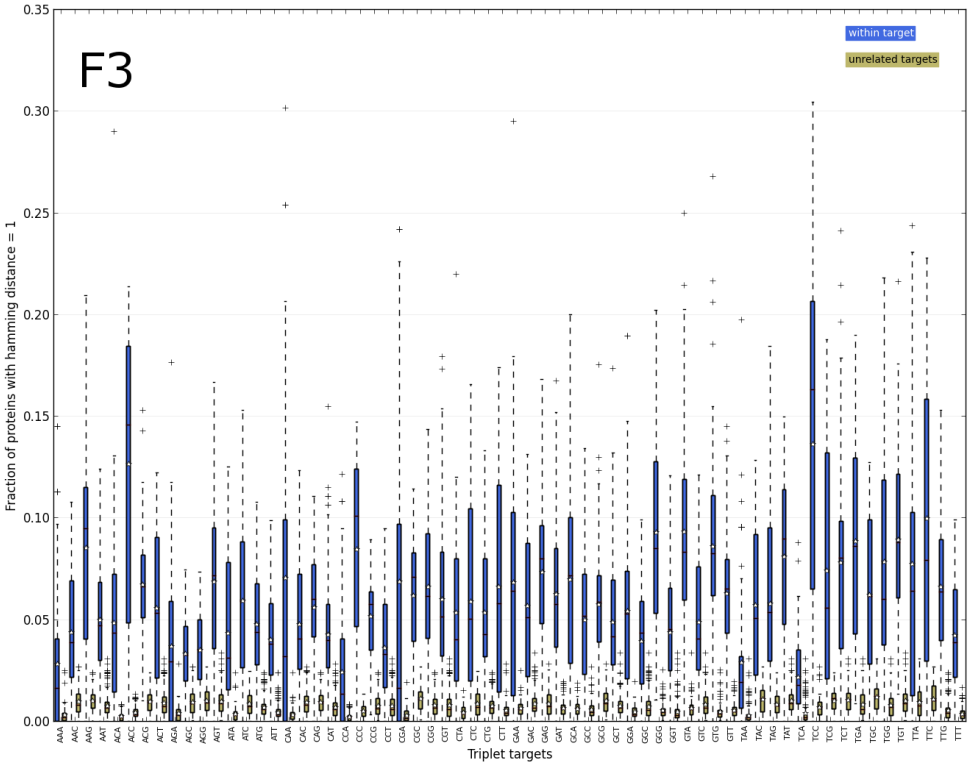
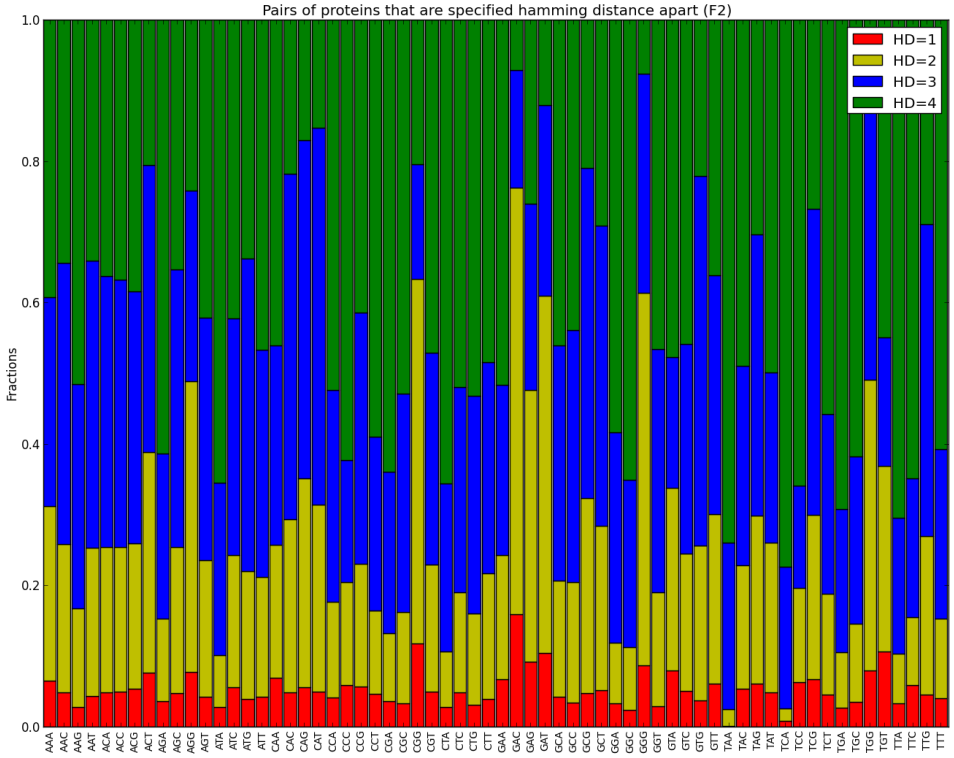

F3
within target
unrelated targets
Fraction of proteins with hamming distance = 1
Triplet targets

Figure S9. Both similar and dissimilar core sequences are found to specify each 3bp target. For each 3bp target, we computed the Hamming distance between all pairs of core sequences assigned to it via lookup using either F2 (top, (a)) or F3 (bottom, (b)) protein selections. For each target in both positions, core sequences that are inferred to specify it show a range of similarities, from pairs differing in only 1 of 4 amino acids in the core positions of the C2H2-ZF domain to pairs differing in all 4 amino acids in the core positions. (Next page.)

A


Pairs of proteins that are specified hamming distance apart (F2)
Fractions
HD=1
HD=2
HD=3
HD=4

B

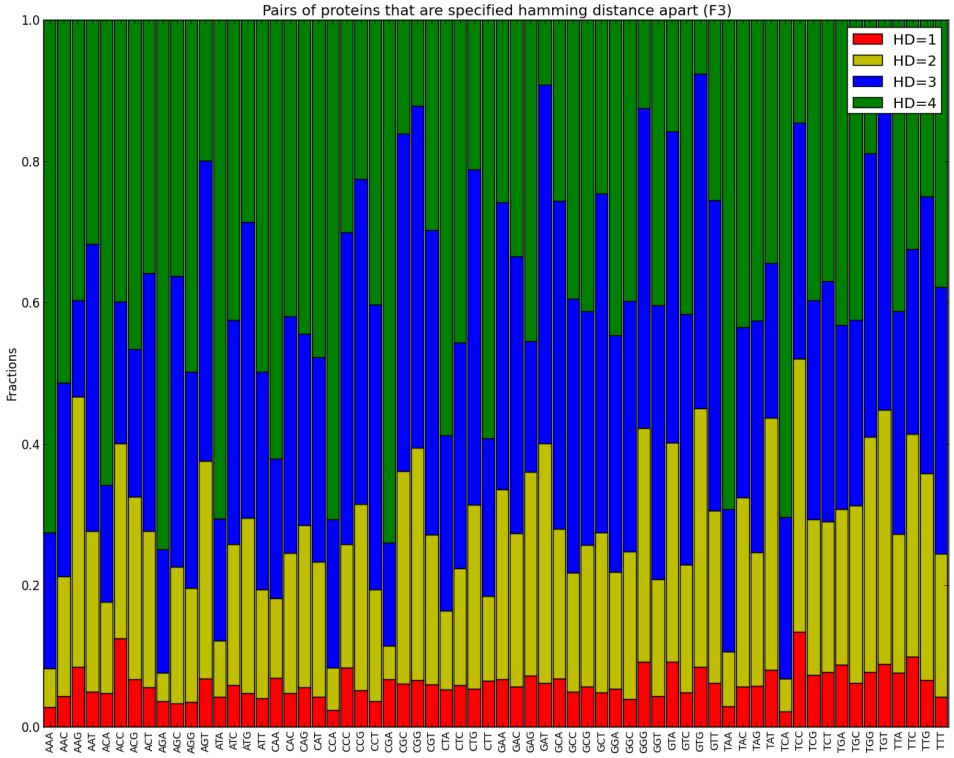

Pairs of proteins that are specified hamming distance apart (F3)
Fractions
HD=1
HD=2
HD=3
HD=4

Figure S10. C2H2-ZF domains with similar and dissimilar sequences selected to bind related targets. Zinc fingers selected to bind common 3bp targets enrich for multiple, dissimilar solutions. Representative candidates tested to bind the related ***A***TG, ***C***TG, and ***G***TG targets are shown. C2H2-ZF domains with dissimilar sequences are able to bind each target as demonstrated by the determined specificities (test helix sequences are noted below each logo, N to C). In fact, 3 or 4 of the core sequence residues may be different, yet the helices still bind a common target such as the ATG and CTG examples shown, though the information content provided may also vary as a result. On the contrary, similar domains can bind different but similar targets where the differences are dictated by a difference in a single contact from the canonical binding model. For example, helices of the sequence trend **TK-aromatic-(T/V)L-*XX*** are tested in each case (boxed column). All of these specify 5' n-TG 3' as expected. The 5' "n" base is dictated by the ***XX*** amino acids that distinguish these domains.

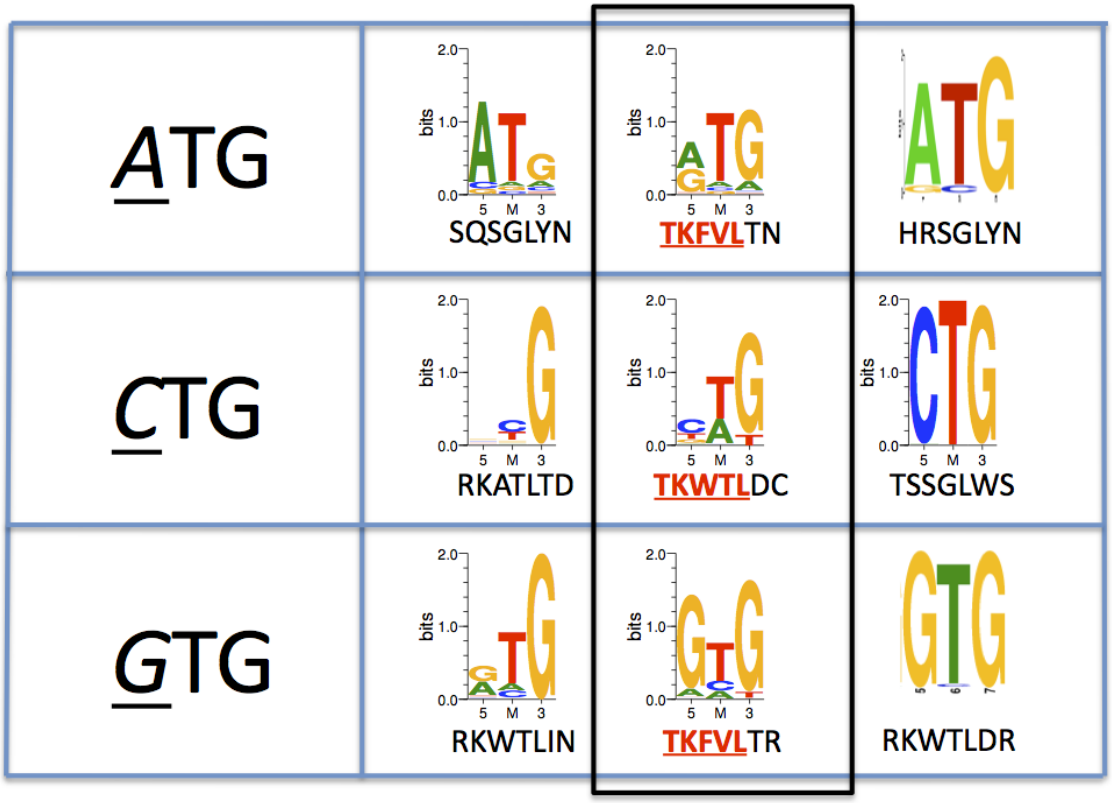


Figure S11. A recreation of the zinc finger code offered in the paper by Wolfe and Pabo noted in the figure. Within each box, the amino acid, or type of amino acid, that is predicted to specify a given base (for a given position of the helix, according to the canonical model) is noted. Core positions of the recognition helix are noted to the left of their specified rows. The predicted base preference is denoted by column, labeled above. The position within the 3-4bp target which is contacted by the specified helical position is shown to the right with an arrow. Additions to the code coming from various literature sources are also noted.

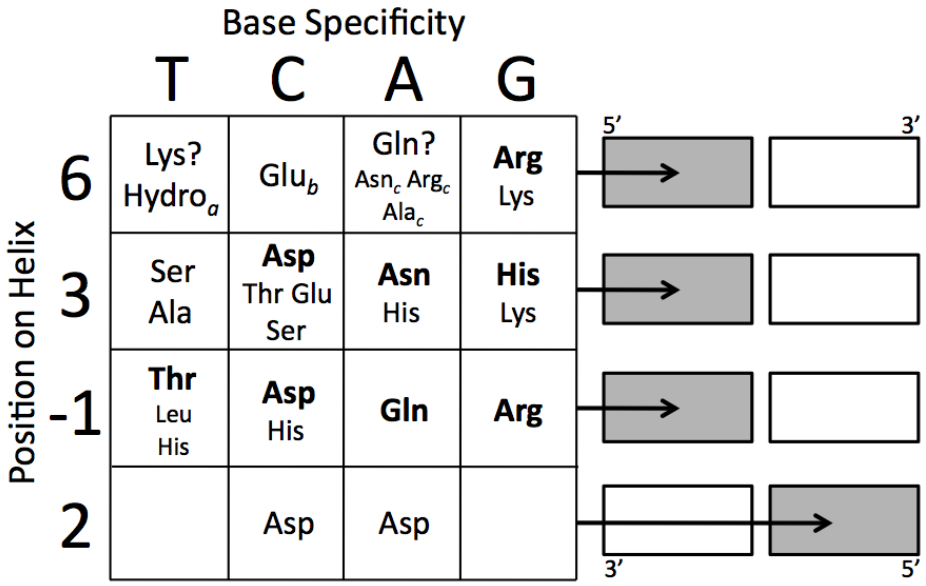


Table modified from:
Wolfe, S.A., Nekludova, L. and Pabo, C.O. (2000) DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev Biophys Biomol Struct*, **29**, 183-212.
Additions:
a Maeder, M.L., Thibodeau-Beganny, S., Osiak, A., Wright, D.A. *et al.* (2008) Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol Cell*, **31**, 294-301.
b Dreier, B., Fuller, R.P., Segal, D.J., Lund, C.V., Blancafort, P., Huber, A., Koksch, B. and Barbas, C.F., 3rd. (2005) Development of zinc finger domains for recognition of the 5'-CNN-3' family DNA sequences and their use in the construction of artificial transcription factors. *J Biol Chem*, **280**, 35588-35597.
c Dreier, B., Beerli, R.R., Segal, D.J., Flippin, J.D. and Barbas, C.F., 3rd. (2001) Development of zinc finger domains for recognition of the 5'-ANN-3' family of DNA sequences and their use in the construction of artificial transcription factors. *J Biol Chem*, **276**, 29466-29478.

Figure S12. Intra-3mer contact at *b3* made by position 2 of the recognition helix. Zinc finger selections reveal common trends between amino acids at position 2 of the helix and the specified 3' base, or as labeled in Figure 1a, "*b3*". (Right) For example, in the F2 protein selections, aromatics at position 2 are enriched for some targets with Guanine at b3. Glutamine is enriched for some targets with Thymine at b3 and Arginine is enriched for targets with Cytosine at b3. Subsequent characterization of DNA-binding specificity confirms these preferences. Representative examples are shown with the position 2 amino acid bold and red in the helical sequence listed below its binding logo. A red arrow points to the b3 position in each column. One exception to the Arg2 trend is shown at the bottom right (VK***R***YLQA) where the Arginine does not lead to b3 Cytosine specificity, which may be influenced by the aromatic at position 3. It should be noted that in all position 2 examples shown there are one or more helices with a Threonine at position -1 while the base preference trends with the position 2 amino acid base preference indicated. (Left, box) A contrary example, where a position 2 amino acid that does not trend with b3 specificity (Serine) is shown. Example Ser2 fingers can show preference for any b3 base. Together, these examples suggest that, positions -1 and 2 influence b3 specificity and a combination of the amino acids at -1 and 2 may contribute to b3 base preference.

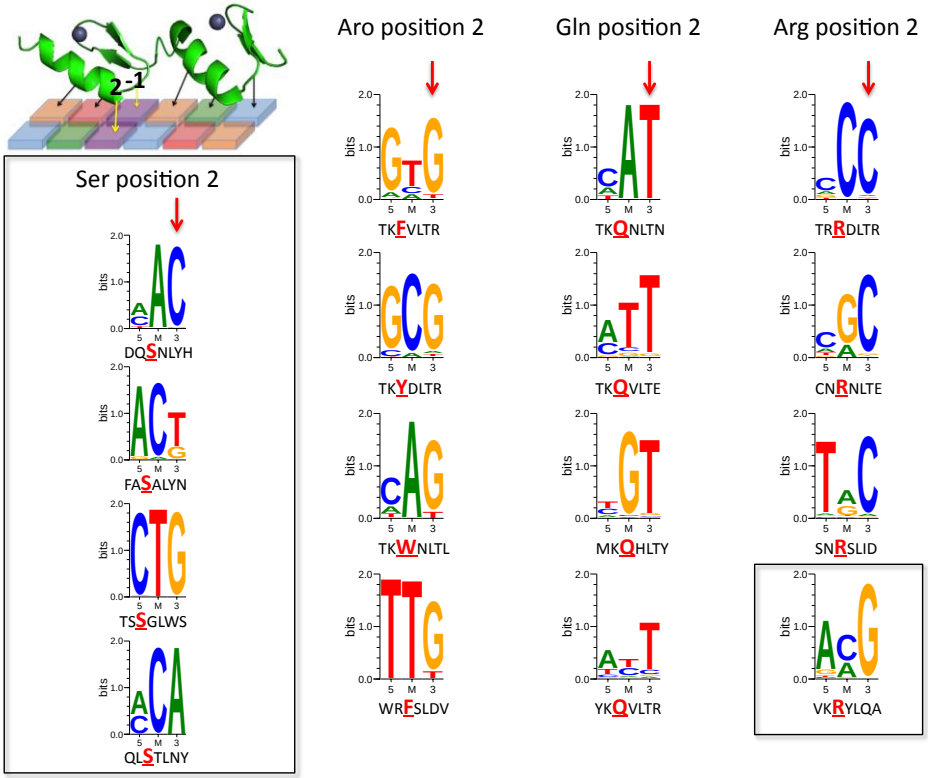


Figure S13. **Core sequences uncovered in both F2 and F3 selections often have dissimilar binding profiles.** We examined the set of core sequences that were selected at both high and low stringency in both F2 and in F3 and that also showed good correspondence with respect to their binding profiles at high and low stringency for each position separately (cosine similarity >= 0.25 in the Figure S6 comparisons). For each core sequence in this set, we compared the binding profile inferred from the F2 protein selections against the binding profile inferred from the F3 protein selections via cosine similarity and plotted the total frequency of core sequences that fell into discrete similarity bins (**Supplemental Methods 2b**). In contrast to the comparisons of low and high stringency data (Figure S5), it is apparent that while much of the density of this histogram still lies in the high cosine similarity bins, a large proportion of it lies in the low to mid-range cosine similarity bins as well. Thus, while there is good correspondence between many binding profiles derived from the F2 vs. F3 contexts, there are also many instances in which the change of positional context leads to starkly different inferred DNA-binding specificities.

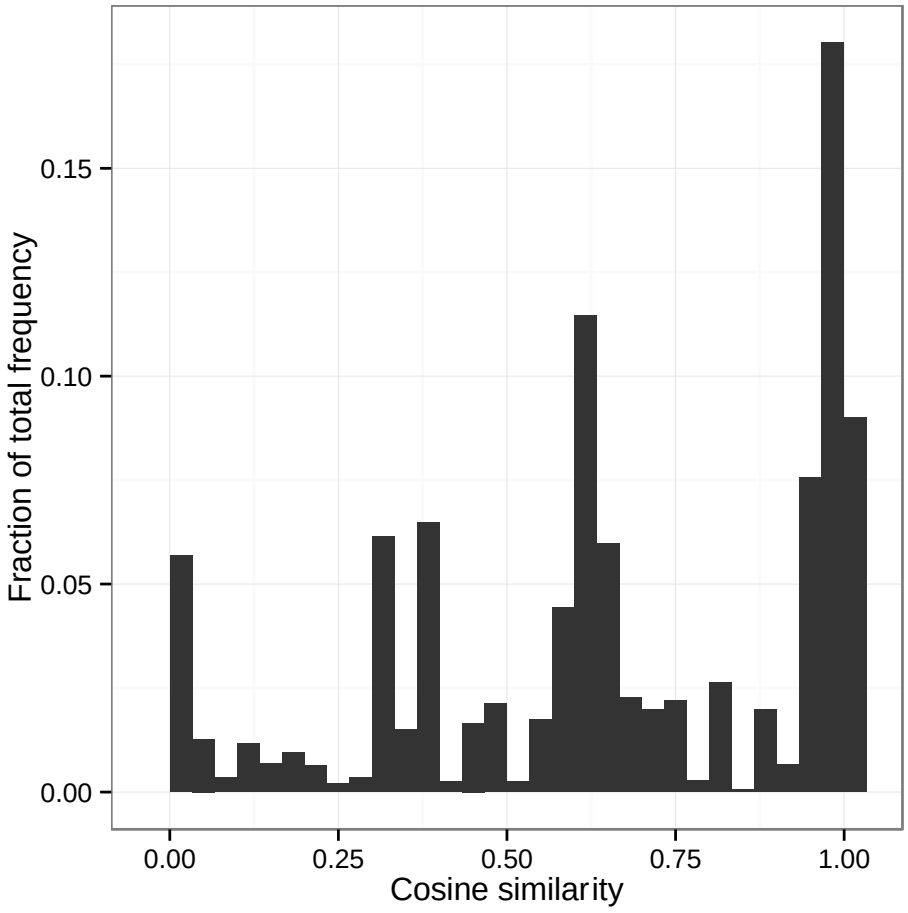


Figure S14. Computationally inferred logos show the least amount of agreement in base position 1 when comparing the F2 vs. F3 context. For each core sequence found to show good agreement between the high and low stringency protein selections within both the F2 and F3 contexts (see Figure S15 caption), we predicted DNA-binding specificities via lookup of the binding profiles (described in **Methods**) using either the F2 or F3 protein selections. For each of these core sequences, the resulting base-frequency distributions for each of the base positions (1, 2, and 3) were compared (F2 vs. F3) using the Pearson correlation coefficient (PCC), and displayed via boxplots. The lowest agreement is observed at base position 1, and these agreements are significantly lower than those observed in base positions 2 and 3 ($p < 0.03$ and $p < 0.0001$, respectively, as computed by the Mann-Whitney U-test). This is notable as this position in the subsite for F3 corresponds to the 5' most base of the entire binding site, and in the subsite for F2 may participate in a cross-strand contact with F3.

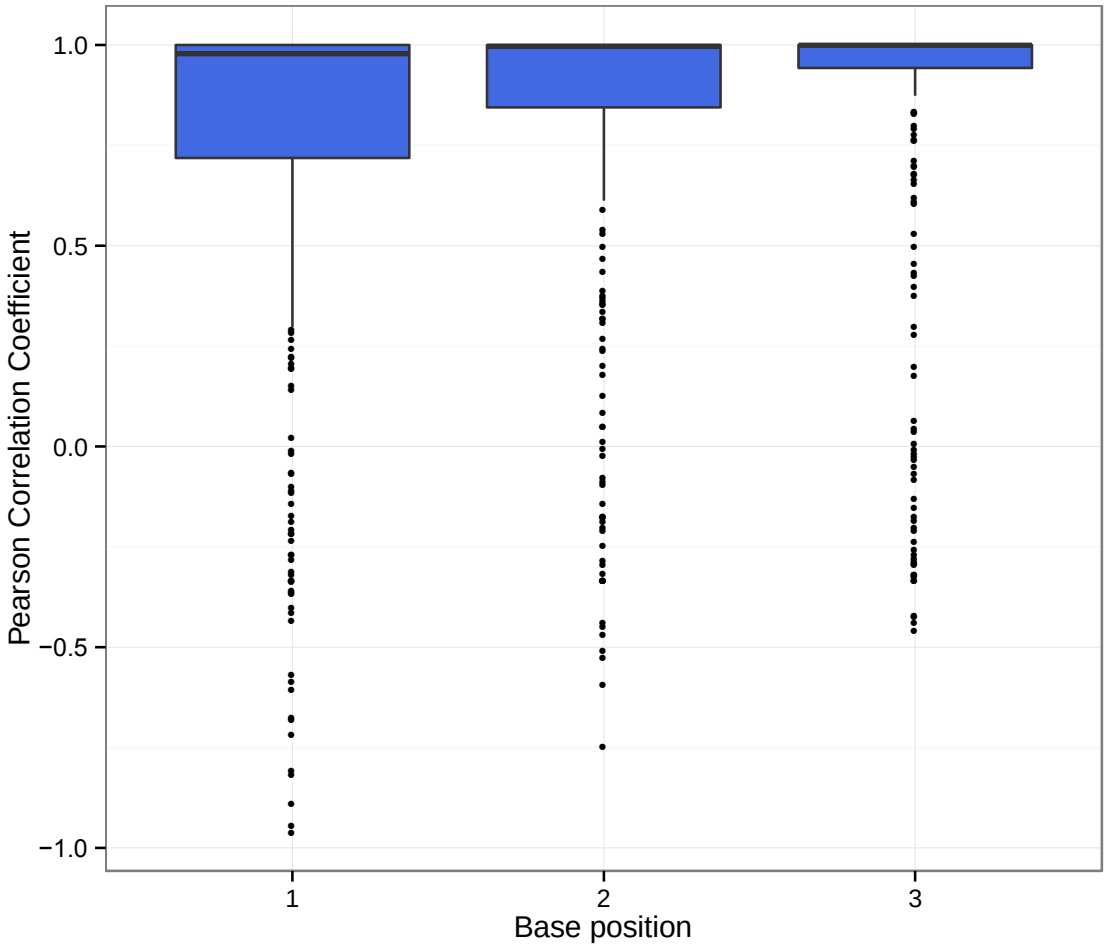


Figure S15. Several of the core sequences tested in both the F2 and F3 positions only exhibit DNA-binding in one of the two positional contexts. (A) Experimental binding site selections were performed in both the F2 and F3 positional context for 26 helices that were identical in the core positions of the recognition helix (-1, 2, 3, and 6). The derived logos show good agreement between the F2 and F3 context for 16 of these helices (top). However, 5 helices showed strong base preferences in the F2 context, but little or no functional DNA-interaction in F3 context (lower left; sequence marked with asterisk produced weak base preferences in the F3 context). Similarly, the remaining 5 helices show good base preferences in the F3 context, but give no evidence of functional interaction in the F2 context (lower right). (B) We produced computationally inferred logos via lookup (see **Methods**) for these same core sequences, based upon either the F2 or F3 protein selection data (F2 logos displayed above F3 logos). These logos show rough agreement between the F2 and F3 context for 13 out of the 16 of the core sequences that showed agreement experimentally (top). Additionally, differential DNA-binding behavior between the F2 and F3 positional contexts was successfully predicted (bottom left and right): in 9 out of 10 cases, no logo is predicted for one of the contexts due to insufficient information in the corresponding protein selection data. This reflects positional biases in protein selections that are analogous to those observed in binding site selections. (Next page.)

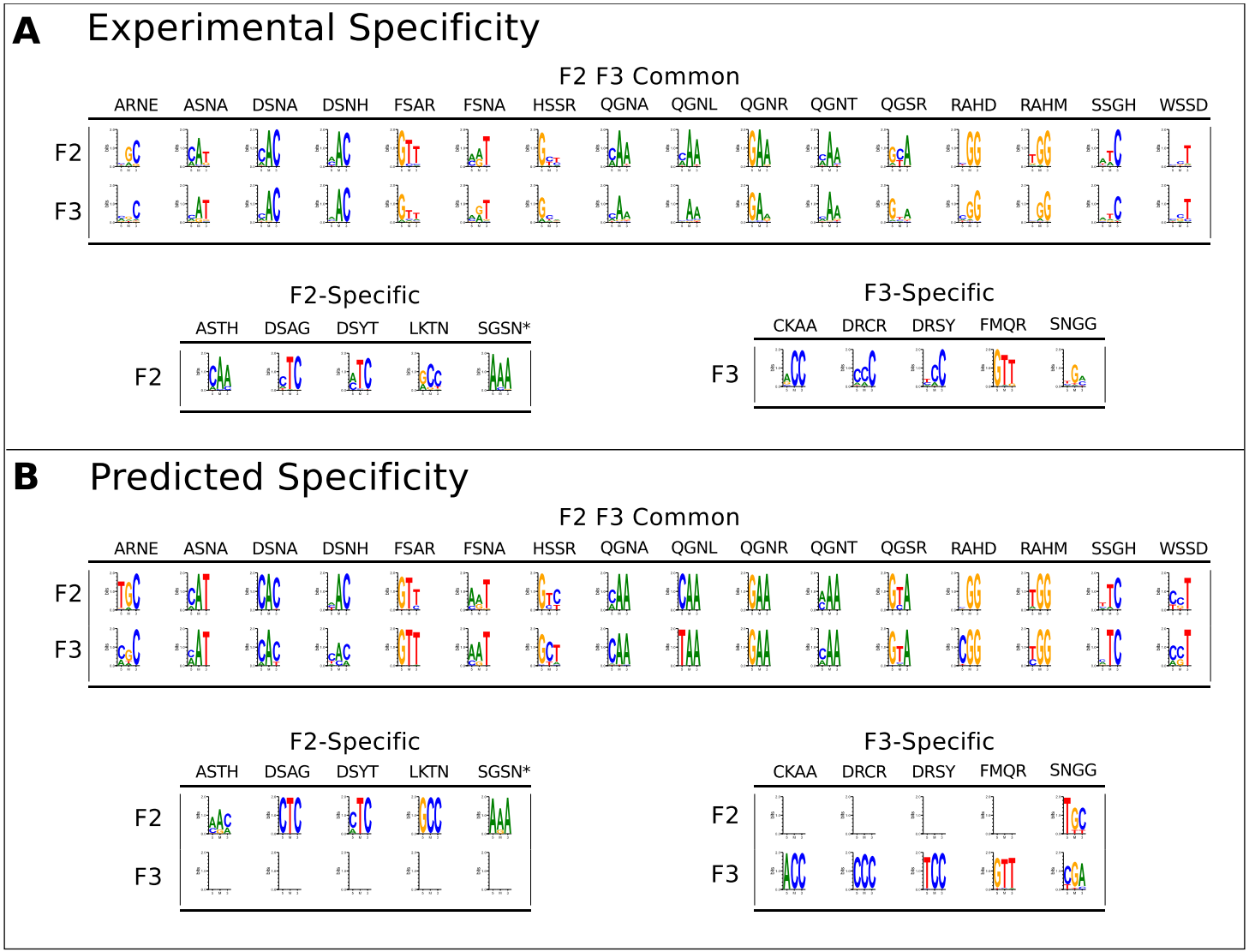

A
Experimental Specificity
F2 F3 Common
ARNE
ASNA
DSNA
DSNH
FSAR
FSNA
HSSR
QGNA
QGNL
QGNR
QGNT
QGSR
RAHD
RAHM
SSGH
WSSD
F2
F3
F2-Specific
ASTH
DSAG
DSYT
LKTN
SGSN*
F2
F3-Specific
CKAA
DRCR
DRSY
FMQR
SNGG
F3
B
Predicted Specificity
F2 F3 Common
ARNE
ASNA
DSNA
DSNH
FSAR
FSNA
HSSR
QGNA
QGNL
QGNR
QGNT
QGSR
RAHD
RAHM
SSGH
WSSD
F2
F3
F2-Specific
ASTH
DSAG
DSYT
LKTN
SGSN*
F2
F3
F3-Specific
CKAA
DRCR
DRSY
FMQR
SNGG
F2
F3

Figure S16. Selected C2H2-ZF assemblies that specify challenging targets.
Three-fingered arrays of C2H2-ZFs were selected from pool-derived libraries to bind targets that arrays constructed by modular assembly failed to recognize. (**Left**) Zinc finger numbers as specified by (Lam et al. 2011) are listed to the left of each row to provide a reference. (**First column**) For each target, the DNA-binding specificity of a selected three-fingered array was determined via B1H DNA-binding site selections and displayed as a sequence logo. Above the logo, the seven amino acids (-1,1,2,3,4,5, and 6) of each helix of the selected three-fingered array are listed N to C, for fingers F3, F2 and F1. Amino acids not coded for in VNS libraries are noted as bold and red. Below each logo, the desired target is provided in colored letters, 5' to 3'. A match between the preferred base of the sequence logo and the desired base in the target is indicated by a capital letter. (**Second column)** The activity of each selected zinc finger array, noted in the first column, was tested in yeast using an affinity-related GFP assay. The zinc finger arrays were challenged to bind the desired target (Correct) or a negative control consisting of an empty vector (Control). Fluorescence was normalized to a positive control and results are displayed as a percentage of this positive control activity.

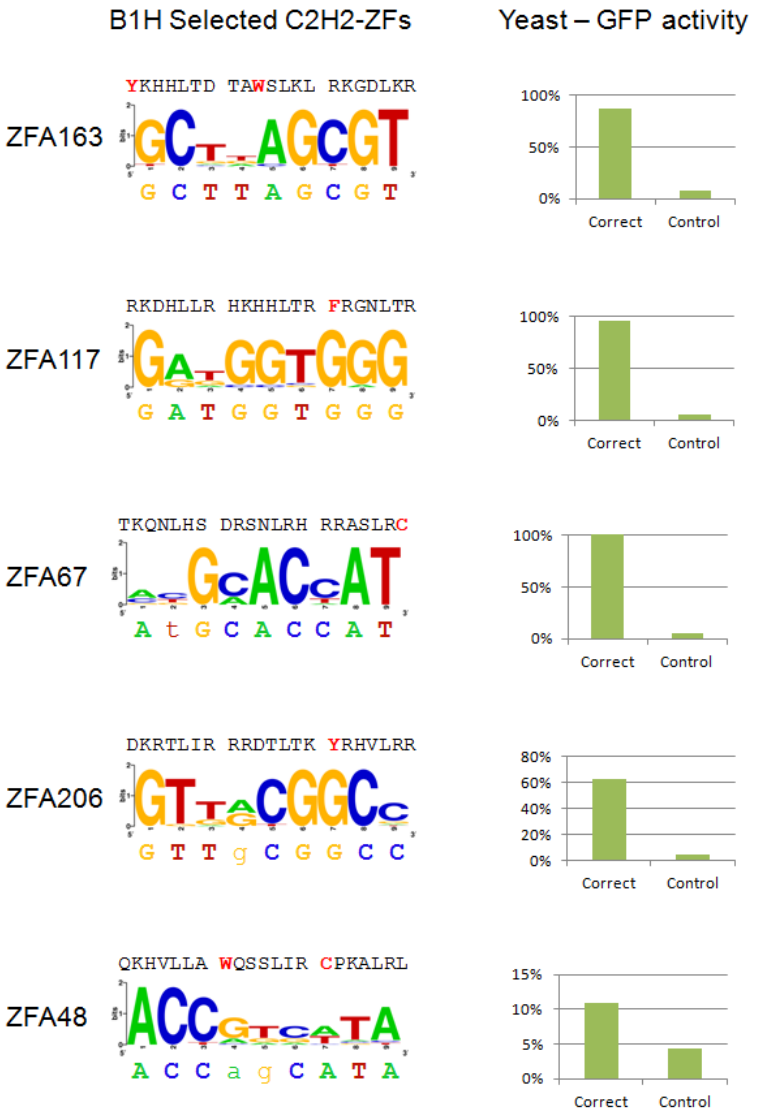
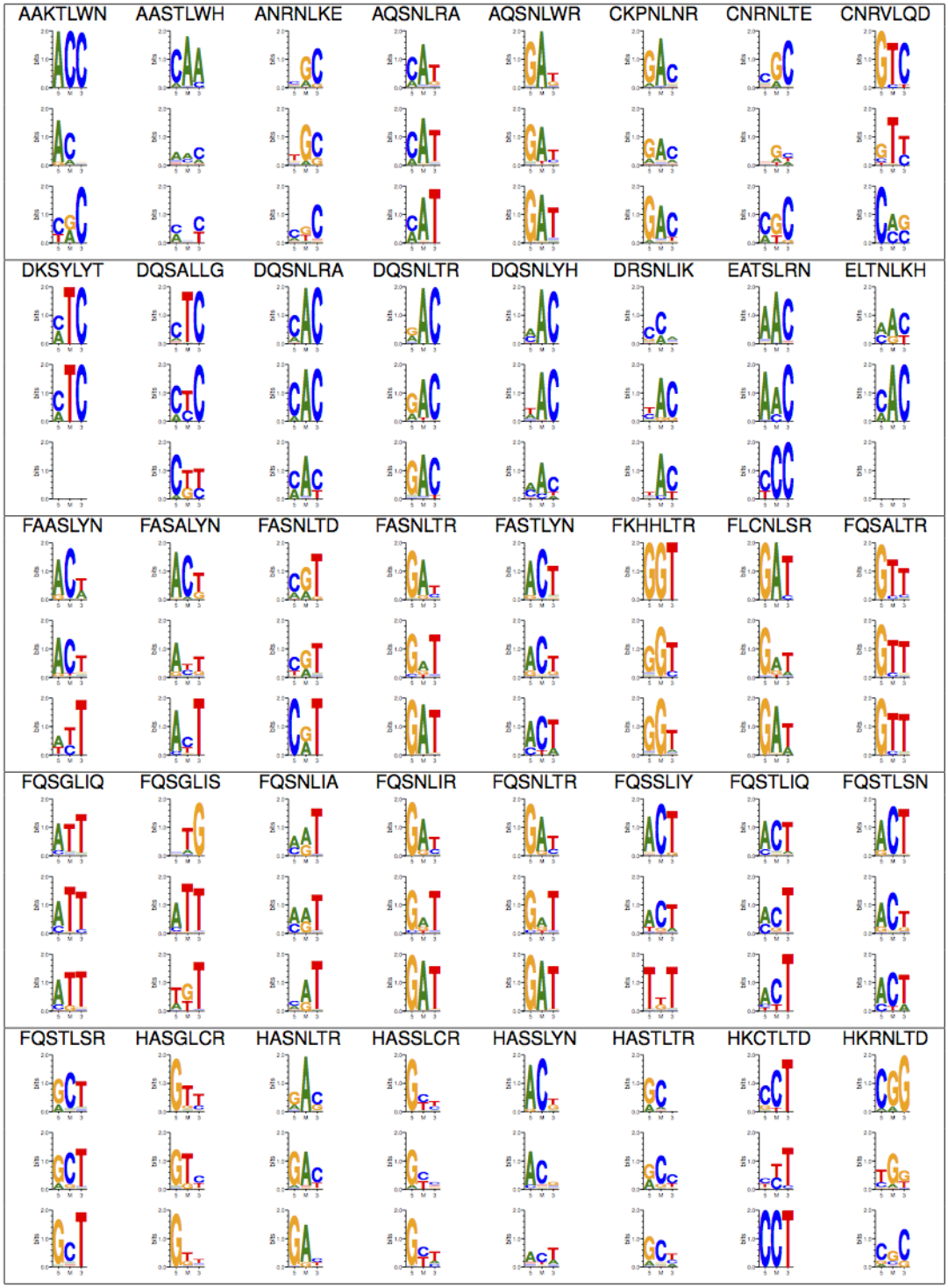
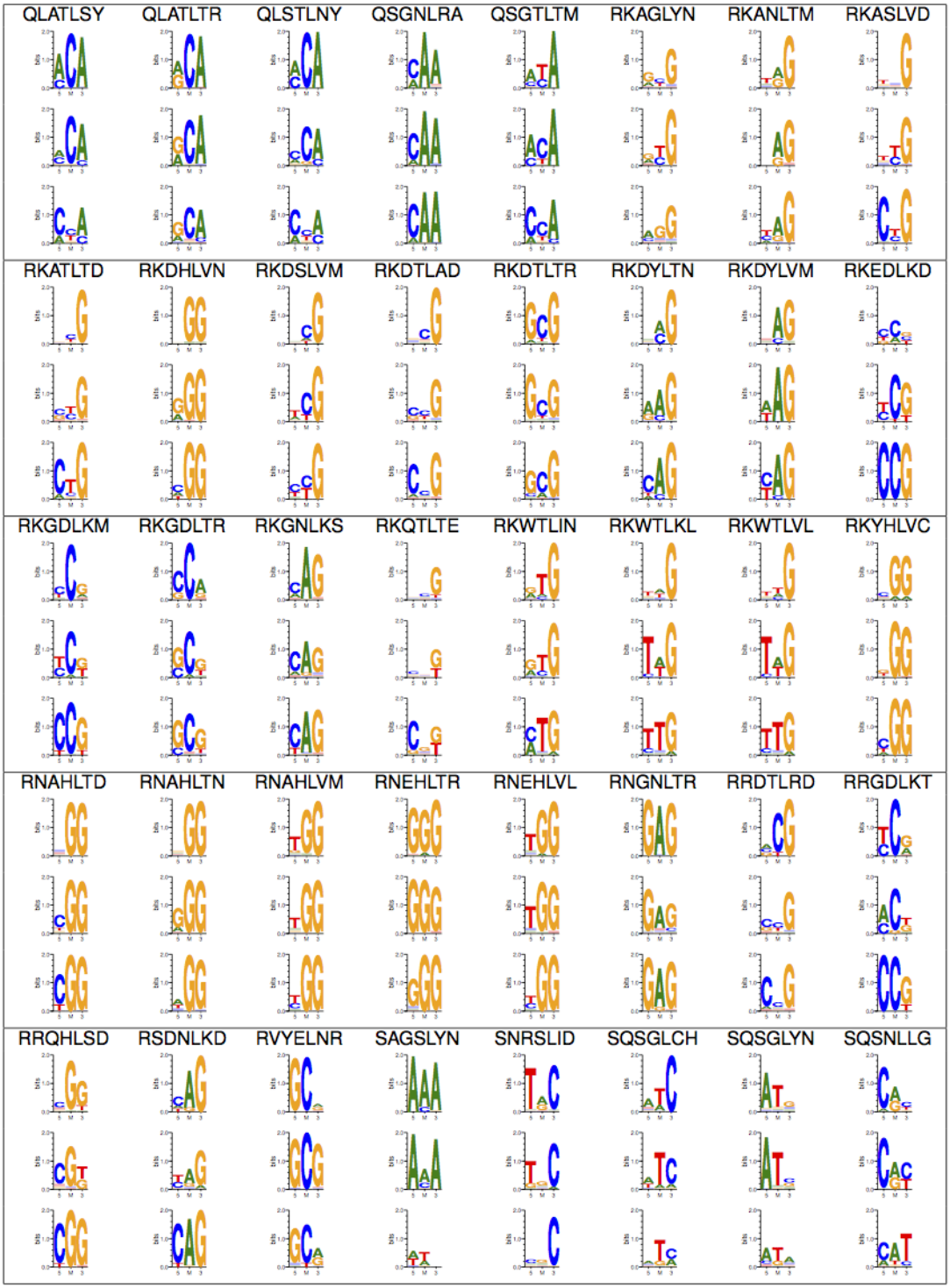


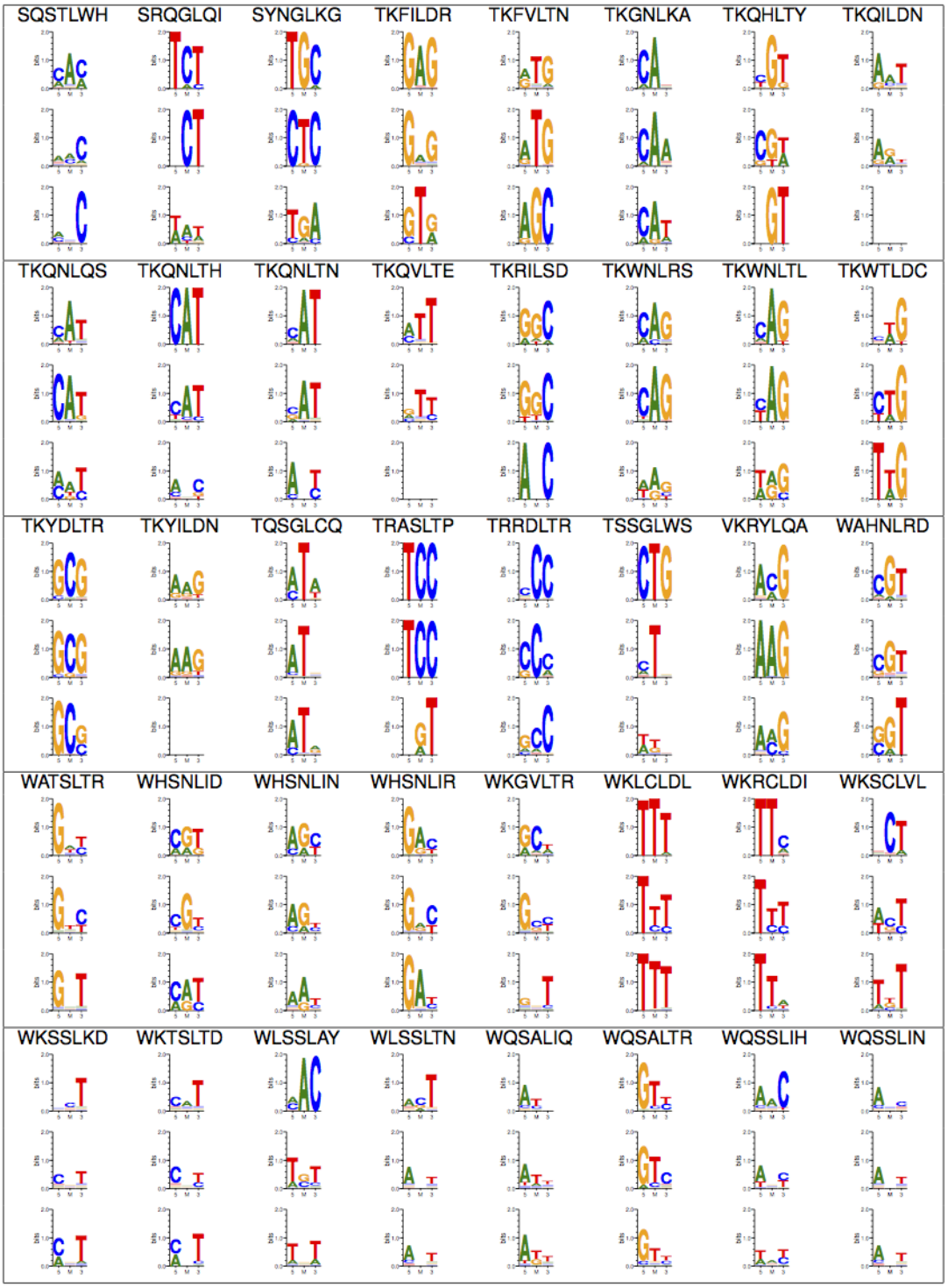
Figure S17. DNA-binding specificities predicted via nearest neighbor decomposition show agreement with experimentally determined DNA-binding specificities. (a) For each of 166 binding site selections performed in the F2 context, we give an experimentally determined logo (top), a logo computationally predicted by nearest neighbor decomposition using the F2 protein selections (middle; ignoring exact matches if present), and a logo computationally predicted by nearest neighbor decomposition using the F3 protein selections (bottom). (b) For each of the 69 binding site selections performed in the F3 context, we give an experimentally determined logo (top), a logo computationally predicted by nearest neighbor decomposition using the F3 protein selections (middle; ignoring exact matches if present), and a logo computationally predicted by nearest neighbor decomposition using the F2 protein selections (bottom). An empty logo corresponds to a case where nearest neighbor decomposition did not predict a specificity, due to no helices in our data set that had the required sequence similarity. (Next 8 pages.)

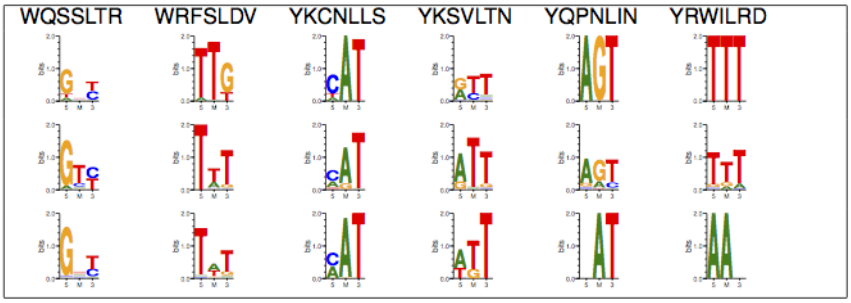
(A)


AAKTLWN
AASTLWH
ANRNLKE
AQSNLRA
AQSNLWR
CKPNLNR
CNRNLTE
CNRVLQD
DKSYLYT
DQSALLG
DQSNLRA
DQSNLTR
DQSNLYH
DRSNLIK
EATSLRN
ELTNLKH
FAASLYN
FASALYN
FASNLTD
FASNLTR
FASTLYN
FKHHLTR
FLCNLSR
FQSALTR
FQSGLIQ
FQSGLIS
FQSNLIA
FQSNLIR
FQSNLTR
FQSSLIY
FQSTLIQ
FQSTLSN
FQSTLSR
HASGLCR
HASNLTR
HASSLCR
HASSLYN
HASTLTR
HKCTLTD
HKRNLTD

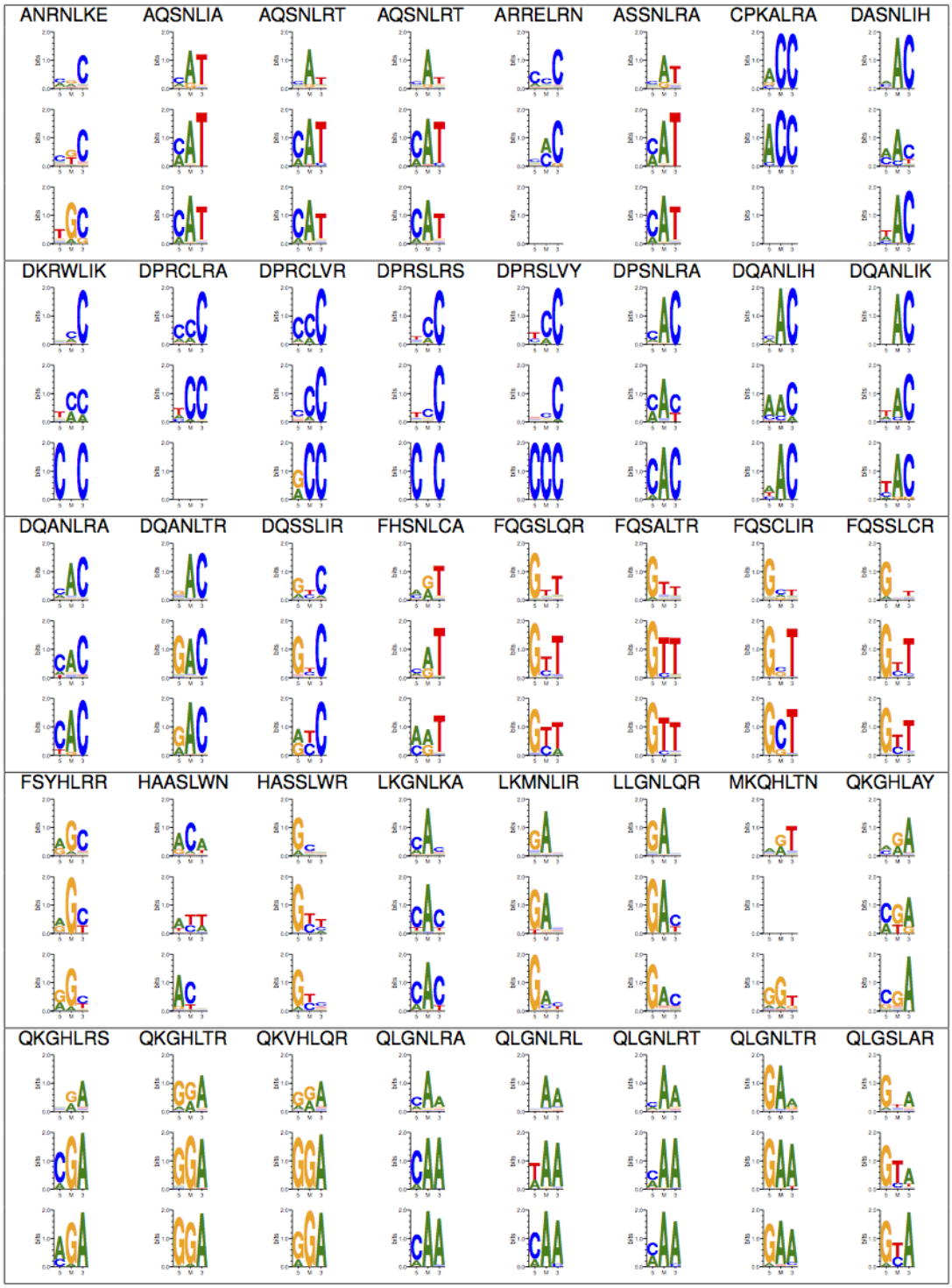

HPSTLGN
HPSTLRY
HPSTLYY
HQANLIH
HQSALYN
HQSGLLG
HQSNLIN
HQSSLCR
HQSSLCR
HQSTLTN
HRSGLYN
HRSNLLG
LKGNLTK
LKKTLIN
LKKTLTD
LNAHLTR
LNDHLQN
MKHHLTD
MKHHLTN
MKHHLTY
MKQHLTN
MKQHLTY
NPGTLWH
NPGTLYY
NSGALWN
QAAGLYN
QDCNLIR
QKAHLTM
QKAHLTR
QKAHLTT
QKGHLST
QKGILTN
QKGNLER
QKGNLKT
QKGNLRL
QKGSLTR
QKINLIN
QKVNLIR
QKVTLTH
QLATLSN


QLATLSY
QLATLTR
QLSTLNY
QSGNLRA
QSGTLTM
RKAGLYN
RKANLTM
RKASLVD
RKATLTD
RKDHLVN
RKDSLVM
RKDTLAD
RKDTLTR
RKDYLTN
RKDYLVM
RKEDLKD
RKGDLKM
RKGDLTR
RKGNLKS
RKQTLTE
RKWTLIN
RKWTLKL
RKWTLVL
RKYHLVC
RNAHLTD
RNAHLTN
RNAHLVM
RNEHLTR
RNEHLVL
RNGNLTR
RRDTLRD
RRGDLKT
RRQHLSD
RSDNLKD
RVYELNR
SAGSLYN
SNRSLID
SQSGLCH
SQSGLYN
SQSNLLG


SQSTLWH
SRQGLQI
SYNGLKG
TKFILDR
TKFVLTN
TKGNLKA
TKQHLTY
TKQILDN
TKQNLQS
TKQNLTH
TKQNLTN
TKQVLTE
TKRILSD
TKWNLRS
TKWNLTL
TKWTLDC
TKYDLTR
TKYILDN
TQSGLCQ
TRASLTP
TRRDLTR
TSSGLWS
VKRYLQA
WAHNLRD
WATSLTR
WHSNLID
WHSNLIN
WHSNLIR
WKGVLTR
WKLCLDL
WKRCLDI
WKSCLVL
WKSSLKD
WKTSLTD
WLSSLAY
WLSSLTN
WQSALIQ
WQSALTR
WQSSLIH
WQSSLIN


WQSSLTR
WRFSLDV
YKCNLLS
YKSVLTN
YQPNLIN
YRWILRD

(B)


ANRNLKE
AQSNLIA
AQSNLRT
AQSNLRT
ARRELRN
ASSNLRA
CPKALRA
DASNLIH
DKRWLIK
DPRCLRA
DPRCLVR
DPRSLRS
DPRSLVY
DPSNLRA
DQANLIH
DQANLIK
DQANLRA
DQANLTR
DQSSLIR
FHSNLCA
FQGSLQR
FQSALTR
FQSCLIR
FQSSLCR
FSYHLRR
HAASLWN
HASSLWR
LKGNLKA
LKMNLIR
LLGNLQR
MKQHLTN
QKGHLAY
QKGHLRS
QKGHLTR
QKVHLQR
QLGNLRA
QLGNLRL
QLGNLRT
QLGNLTR
QLGSLAR

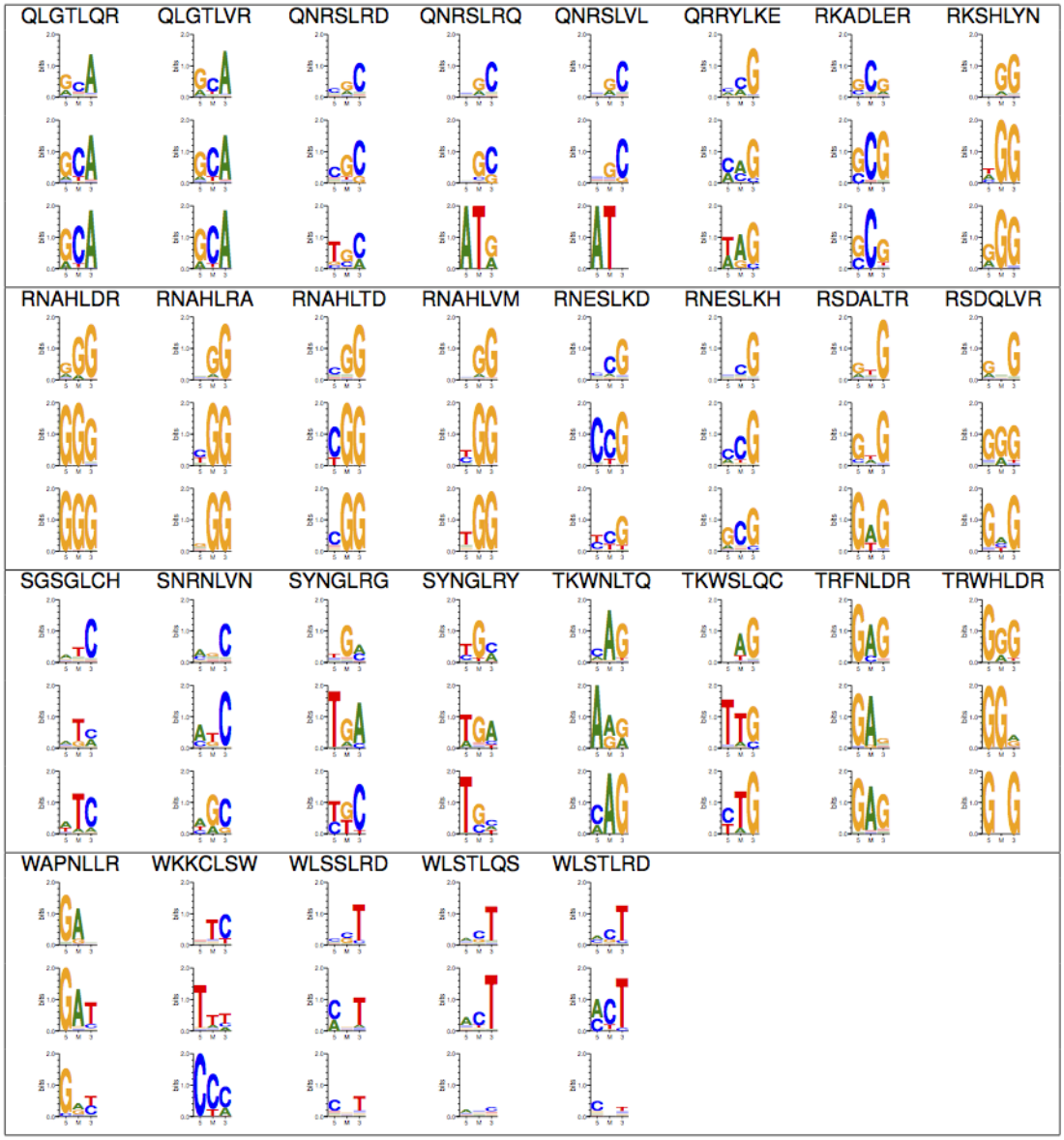

QLGTLQR
QLGTLVR
QNRSLRD
QNRSLRQ
QNRSLVL
QRRYLKE
RKADLER
RKSHLYN
RNAHLDR
RNAHLRA
RNAHLTD
RNAHLVM
RNESLKD
RNESLKH
RSDALTR
RSDQLVR
SGSGLCH
SNRNLVN
SYNGLRG
SYNGLRY
TKWNLTQ
TKWSLQC
TRFNLDR
TRWHLDR
WAPNLLR
WKKCLSW
WLSSLRD
WLSTLQS
WLSTLRD

Figure S18. Predictions based on nearest neighbor decomposition (NN) effectively extend the data to account for differences in positional context. The fraction of predictions with >= 3, 2, 1 or 0 columns correct (as judged by a PCC >= 0.5) is shown for the NN approach (described in **Methods**) and compared to two previous prediction approaches based on random forests (RF) or support-vector machines (SVM). (**A**) Results for the 166 C2H2-ZFs that were tested as F2 in our B1H DNA-binding site selections. NN predictions were computed based on the F3 protein selections. (**B**) Results for the 69 C2H2-ZFs that were tested as F3 in our B1H DNA-binding site selections. NN predictions were computed based on the F2 protein selections.

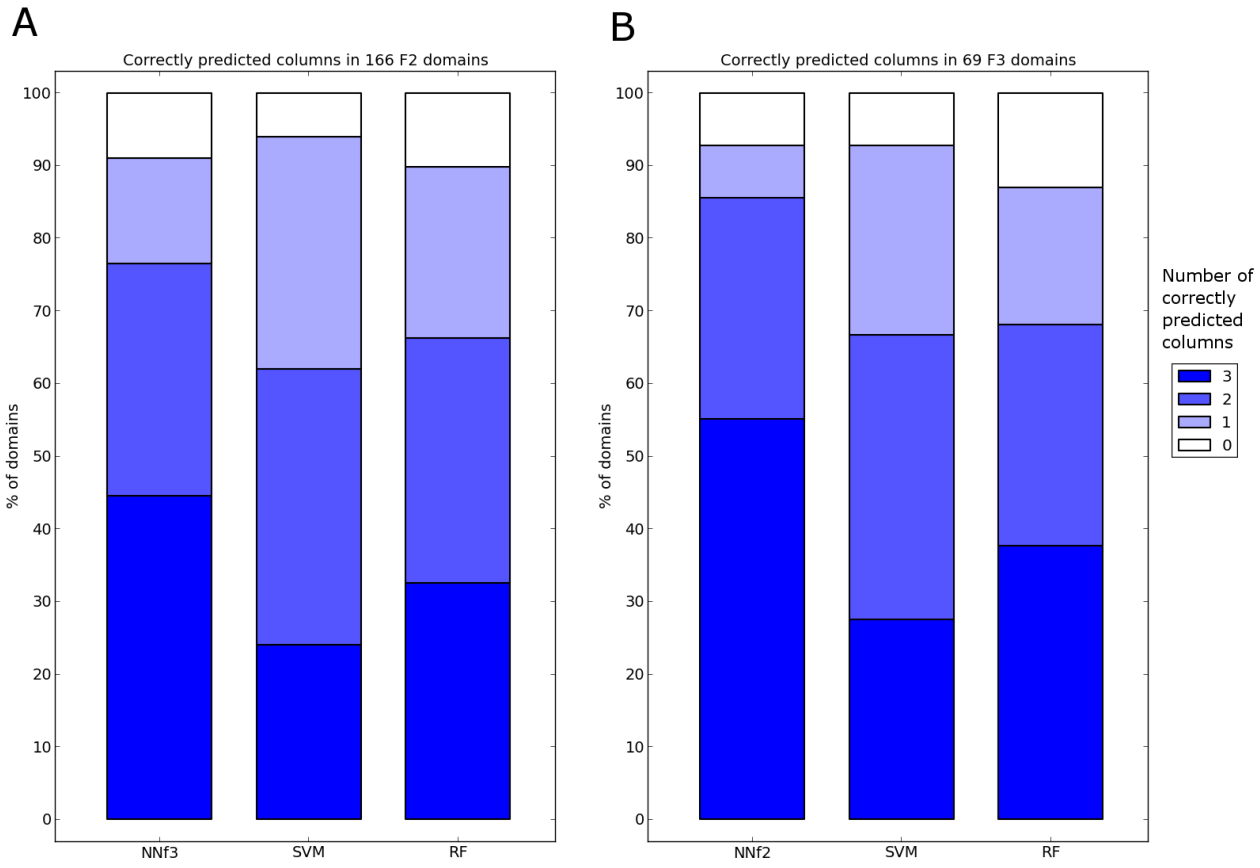


Figure S19. Predictions based on nearest neighbor decomposition outperform the lookup approach that is based solely a single core sequence. We compare the performance of our original lookup inference approach (based solely on the binding profile of the exact core sequence) with the performance of our nearest neighbor decomposition approach (both approaches described in **Methods**). (A, left) Fraction of correctly predicted per-nucleotide base preferences, as judged by a PCC >= 0.5, on the 166 helices tested in the F2 positional context based upon either F2 or F3 protein selections using either lookup or nearest neighbor decomposition (ignoring exact matches for predictions based on F2 protein selection data). (A, right) Fraction of these predicted 3bp binding specificities that have 0, 1, 2, or 3 base preferences correctly predicted. (B, left) Fraction of correctly predicted per-nucleotide base preferences, as judged by a PCC >= 0.5, on the 69 helices tested in the F3 position context based upon either F2 or F3 protein selections using either lookup or nearest neighbor decomposition (ignoring exact matches for predictions based on F3 protein selections data). (B, right) Fraction of these predicted 3bp binding specificities that have 0, 1, 2, or 3 base preferences correctly predicted.

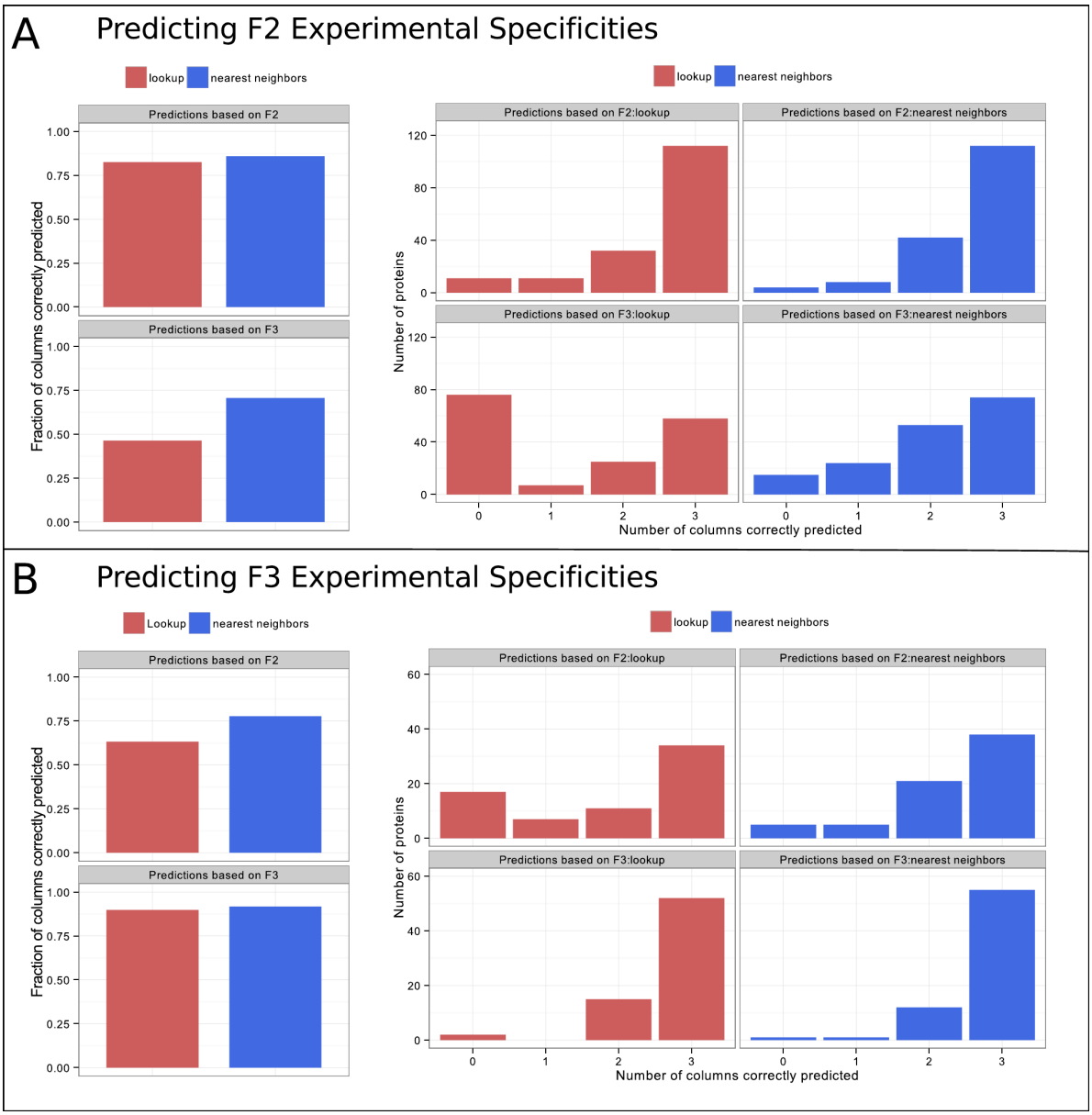


Figure S20. The nearest neighbor decomposition method (NN) performs comparably to other state-of-the-art C2H2-ZF binding site prediction methods (SVM and RF) for predicting binding sites of naturally occurring C2H2-ZF arrays. (A) The percent of proteins (y-axis) with statistically significant alignments between the predicted and experimental PWMs (Supplemental Methods 2e), shown across various p-value thresholds (x-axis). (B) The percent of proteins (y-axis) whose aligned predicted and experimental PWMs report at least a given percent of columns correctly predicted (x-axis), using an PCC threshold of 0.5 to determine whether an aligned column is correct. For both panels, NN is shown in green, RF in black and SVM in blue.